A FRAMEWORK FOR DEMONSTRATING PRACTICE SCHEDULE EFFECTS IN SKILL ACQUISITION

A Thesis
Presented to
The Academic Faculty

by

Brian D. Gane

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology
December 2011

A FRAMEWORK FOR DEMONSTRATING PRACTICE SCHEDULE EFFECTS IN SKILL ACQUISITION

Approved by:

Professor Richard Catrambone,
Committee Chair
School of Psychology
Georgia Institute of Technology

Professor Arthur D. Fisk
School of Psychology
Georgia Institute of Technology

Professor Wendy A. Rogers
School of Psychology
Georgia Institute of Technology

Professor Eric Schumacher
School of Psychology
Georgia Institute of Technology

Professor David Washburn
School of Psychology
Georgia Institute of Technology

Date Approved: 27 October 2011

ACKNOWLEDGEMENTS

This research would not have been possible with out the help of many individuals throughout the years. Helen Alvey, Chance Hughes, and Aleks Petkov spent countless hours running participants, coding data, and assisting me with pilot testing and development. Brian Lawrence, Lina Lee, Zach Malone, George Mitchell, and James Oguntebi authored the software to draw the stimuli for Experiments 3 and 4. Without their efforts, and in particular, without the extra work of Zach and James, those experiments would not have gotten off the ground. Thanks to Brian Bohannon and Tom Reddick who provided expertise during early development of the task used in Experiments 3 and 4.

This research was partially supported by a Georgia Tech School of Psychology Graduate Award. In addition, CEISMC provided a stable source of funding during my last two years.

The solid theoretical and methodological aspects of this research are indebted to the discussions and feedback from Richard Catrambone and the graduate students in the PSET Lab. The gaps in theory and methodology are likely a result of me ignoring their advice. Thanks to my dissertation committee for pushing me to turn a loose set of ideas into a solid set of empirical studies.

Finally, this research would not have been possible without the loving support of Megan and my family, who provided encouragement and motivation when I needed it most. Many thanks to Kira and Chance for rigidly enforcing study breaks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

I outline a framework for researching the effects of practice schedule on skill acquisition, based upon stage theories of information processing and stage theories of skill acquisition. Skilled performance requires stimulus identification, response selection, and response execution. I hypothesize that practice schedule affects learning in two types of information processing stages: stimulus-oriented and response-oriented stages. The loci of these effects differ based on the stage. In stimulus-oriented stages, practice schedule affects concept and categorization learning via contiguity of exemplars and feature saliency. In response-oriented stages, practice schedule affects the efficiency with which individuals produce a response by affecting response preparation. I evaluated this framework and theory with 4 experiments that manipulated practice schedule and amount of practice, in 2 domains with different information processing demands. Experiments 1 and 2 focused on response-oriented stages via a task that required participants to execute a multisegment movement according to a target time. Experiments 3 and 4 focused on stimulus-oriented stages via a task that required participants to categorize football play diagrams. Within the 2 task domains the amount of acquisition practice was manipulated to test whether different durations of acquisition training changed how practice schedules affected retention and transfer performance. The practice schedule manipulation had reliable effects on performance and learning when task performance involved either response preparation or induction of categorization rules. Practice schedule did not affect performance or learning when task performance involved categorization decisions, after the rules had been learned. Additionally, I report a novel method for quantifying amount of practice that allows comparisons across task domains.

CHAPTER I

INTRODUCTION

When designing a training or educational program, sequencing instruction is an important step. One must decide which information a learner should see first, when new information should be introduced, and how to schedule the presentation of multiple pieces of information. Each of these decisions can make a difference in learning, thus one must be aware of how *order effects* can potentially affect learning and training environments (Ritter, Nerb, Lehtinen, & O'Shea, 2007).

Consider domains involving psychomotor, perceptual, and cognitive skills. For instance, consider a flight simulator training program, a course for radiologists learning to identify different anatomical structures using different imaging techniques, or a classroom environment, such as physics, in which learners must learn to solve a variety of related problems: conservation problems, momentum problems, etc. In each of these examples, the instructional designer must decide whether these different tasks should be taught sequentially, or in a mixed fashion. For instance, should a flight simulator vary whether trainees land from a different approach, in different weather conditions, and at a different airport on each attempt? Should a radiologist study multiple different images of the same structure, all in the same session? When designing a physics textbook, should the exercises be grouped such that learners solve a series of problems from the same category? This decision of how to sequence instruction becomes more challenging for trainers and educators attempting to assess effectiveness of the educational program, given that immediate performance might not reflect performance after a delay or transfer to a new situation (Schmidt & Bjork, 1992; W. Schneider, 1985).

One goal of this dissertation is to provide a framework for investigating order effects in the skill acquisition literature. I present a task-analytic framework that is not tied to a specific skill domain, and so, it is hoped, will be useful in future research investigating order effects in training psychomotor, perceptual, or cognitive skills. One specific type of order effect, *practice schedule*, will be investigated, and will be linked to a primary research area in the skill acquisition literature, the *contextual interference effect*. Additionally, the practice schedule variable will be manipulated under different durations of acquisition training to characterize how the *amount of practice* might moderate the effects of practice schedule. Before introducing this research framework I review the general state of practice schedule research in the skill acquisition literature.

1.1 Practice Schedule Research in Skill Acquisition

In the skill acquisition literature, one order effect that has generated much experimental study is practice schedule, the scheduling of practice on multiple tasks. The explanation most frequently used to frame the effects of practice schedule on skill acquisition is the contextual interference effect.

Contextual interference (CI) occurs when the context in which a task is practiced creates interference. The *contextual interference effect* refers to situations in which this interference alters the cognitive processing that occurs during acquisition. As a result of this altered processing, learning is affected, as measured by acquisition, retention, or transfer performance (Battig, 1979; Magill & Hall, 1990; Schmidt & Bjork, 1992).

A variety of contexts, and thus sources of interference, are possible. The contextual interference research literature, however, has most often focused on manipulating *contextual variety*, by manipulating the intervening *tasks* on trial$_{n-1}$ and trial$_{n+1}$, etc. (e.g., Carlson & Yaure, 1990; J. B. Shea & Morgan, 1979).

In a typical experiment manipulating practice schedule in skill acquisition (e.g., J. B. Shea & Morgan, 1979), a skill domain with multiple tasks is selected (e.g., learning to solve four types of physics word problems). Practice schedule is manipulated. Training follows either a *blocked schedule*, in which participants practice a series of examples sampled from one task before moving on to the next task, or a *random schedule* in which participants practice a series of examples sampled from the entire set of examples, with no blocking of examples by task type. In the typical CI effect, acquisition performance favors the blocked schedule (e.g., speeded movements are faster in the blocked schedule condition), whereas retention and transfer performance typically favor the random schedule (e.g., speeded movements are faster in the random schedule condition; Magill & Hall, 1990). To be clear, practice schedule refers to an instructional design feature, whereas contextual interference refers to a psychological theory as to why practice schedule creates the observed pattern of superior acquisition performance and inferior retention and transfer performance for the low contextual interference group (i.e., the blocked practice schedule group). Contextual interference has been studied in a variety of domains, such as multisegment movement tasks (e.g., J. B. Shea & Morgan, 1979), knot-tying tasks (Ollis, Button, & Fairweather, 2005), ATM menu navigation tasks (Jamieson & Rogers, 2000), Boolean logic rules (e.g., Carlson & Yaure, 1990), and foreign language acquisition (V. I. Schneider, Healy, & Bourne Jr., 2002).

1.1.1 Theories of Contextual Interference

The literature on contextual interference contains two general, high-level explanations for why a random schedule results in more learning than a blocked practice schedule (assuming a sufficient amount of practice). The two explanations are the *elaborative processing view* (J. B. Shea & Morgan, 1979; J. B. Shea & Zimny, 1983) and the *reconstruction view* (Lee & Magill, 1983).

1.1.1.1 Elaborative Processing View

In the elaborative processing view multiple, variable processing during acquisition improves retention (J. B. Shea & Morgan, 1979). This could be either intraitem or interitem processing. According to J. B. Shea and Zimny (1983), in a random schedule (high CI), multiple tasks will be held in working memory simultaneously and thus allow interitem processing to occur. In a blocked schedule (low CI), only a single task is held in working memory and thus only intraitem processing can occur. J. B. Shea and Zimny (1983) argue that variable, multiple processing is important, and thus the random practice schedule is beneficial because participants can use *both* interitem and intraitem processing.

1.1.1.2 Reconstruction View

In the reconstruction view (Lee & Magill, 1983), high CI causes participants to reconstruct their *action plan* before each trial. This reconstruction is effortful which decreases performance during acquisition. The result of this trial by trial reconstruction, however, is that individuals have a greater ability to retrieve those action plans during retention trials, and thus have higher performance. Relatedly, Carlson has argued that random trials lead to more practice loading operators into working memory (comparable to retrieving action plans), and that this retrieval from long term memory is effortful and slows acquisition performance, but ultimately improves performance on transfer tests and tests that require sequencing overlapping operations in working memory (Carlson & Shin, 1996; Carlson & Yaure, 1990).

I argue that these contextual interference theories can be further refined by including the components of skilled performance that are affected, and how this changes with different amounts of acquisition practice. A component-based approach to CI effects offers a perspective to integrate these two existing theories: both types of processing mechanisms (elaboration and reconstruction) can be affected by practice schedules, but which mechanism is affected depends on the component of the task that most strongly drives successful performance. Specifically, the elaborative processing account is most relevant to contextual interference effects in stimulus learning and the reconstruction view is most relevant to contextual interference effects in response learning.

1.1.2 Contextual Interference in Different Skill Domains

Although the term contextual interference originates from research in verbal learning (Battig, 1979), by far the most consistent results supporting the presence of contextual interference effects occur in motor learning research. The demonstration that has generated the most research in motor learning was first reported by J. B. Shea and Morgan (1979), who manipulated contextual variety in a barrier-knockdown task[1]. Since J. B. Shea and Morgan, many studies have replicated this basic effect. Two major reviews of CI in motor skill learning show that CI aids performance in skills when the tasks utilize different motor programs (Magill & Hall, 1990; see also Lee, Wulf, & Schmidt, 1992), and that the evidence for CI is stronger in simple motor skills than in complex motor skills (Wulf & Shea, 2002). Together, the reviews suggest simple motor skills (e.g., a multi-segment movement task) that require different planning operations seem to show the strongest effects of CI.

Unlike the strong evidence of CI in motor learning, the evidence in cognitive skill learning is more sparse. The initial demonstrations of CI were in the verbal learning literature (Battig, 1972, 1979; Pellegrino, 1972). The promising results in the motor literature (J. B. Shea & Morgan, 1979; Lee & Magill, 1983) then motivated new attempts to extend the CI effect to skills that are predominately cognitive. Notably,

[1]Participants had to knock over a series of blocks arranged in a 2 x 3 grid. During acquisition, participants performed three barrier-knockdown tasks, each requiring a different order and subset of barriers to knock down. In the blocked schedule condition, participants completed all 18 practice trials for a task before completing all 18 practice trials for the next task. In the random schedule condition, the practice trials were ordered such that participants never completed more than two practice trials for each task in succession.

Carlson and colleagues (Carlson & Shin, 1996; Carlson & Yaure, 1990) have shown impressive results by manipulating the practice schedule in which participants learn to solve Boolean logic gate problems. More recently, Bjork and colleagues reported results showing that a random schedule can increase accuracy in a category learning task (Kornell & Bjork, 2008). Unfortunately, other attempts to extend the effect to other cognitive tasks have yielded equivocal results, including weak effects or unclear effects (Jamieson & Rogers, 2000; V. I. Schneider et al., 2002; van Merriënboer, Shuurman, de Croock, & Paas, 2002), no effects (Quilici & Mayer, 1996, Experiment 3; van Gerven, Paas, van Merriënboer, & Schmidt, 2006), and reverse effects (Gane & Catrambone, 2009).

These results of practice schedule effects in skill acquisition might seem to suggest that CI has an important role in simple motor skills, but possibly only a limited role in cognitive skills and complex motor skills. However, echoing previous skill researchers (e.g., Fitts, 1964), I argue that dichotomizing these skill acquisition findings based on domain is incorrect. Instead, based on a review of the literature, I submit that one critical variable that characterizes the differences between replications and non-replications of CI is the component of performance that the order manipulation affects. This view extends previous researchers' call for examining the components of the task that enable performance in order to see what training components should be manipulated with practice schedule (Lee et al., 1992).

1.2 Stages of Information Processing

In general, theories of skilled performance propose different stages of information processing. For instance, researchers acknowledge a variety of stages (stimulus encoding, stimulus categorization, response selection, response planning, response programming, and response execution stages), though they differ in which subset of stages they emphasize (Schmidt & Lee, 2005). These stages might not be completely independent, but nonetheless are reasonably separable and can act as a theoretical means to divide the time course and processes involved in information processing.

1.3 Stages of Skill Acquisition

Additionally, a variety of theories of skill acquisition propose different stages of skill learning. One class of skill acquisition theories (Charney & Reder, 1986; McGuire, 1961; E. E. Smith, Chase, & Smith, 1973; Underwood & Schulz, 1960) share in common a general division: stimulus learning (concept learning, stimulus categorization, encoding), response learning (response integration, executing procedures, response execution), and associative learning (S-R mappings, learning when to execute procedures). Additionally, another class of skill acquisition theories adopts a perspective focused on qualitative processing changes over the course of skill acquisition: cognitive, associative, autonomous (Fitts & Posner, 1967), and encoding, procedural, and compilation (Neves & Anderson, 1981). These two classes of stage theories map onto each other. For instance, Underwood's association stage maps onto Anderson's encoding stage and Underwood's response learning stage maps onto Anderson's proceduralization and compilation stages. Both classes of stage theories use stages that are theoretically distinct, but might not be cleanly separable in actual skilled performance (e.g., McGuire, 1961; E. E. Smith et al., 1973). At any given time, components of a complex skill might be at different stages of acquisition as learners do not cleanly shift from one stage to another (Fitts & Posner, 1967; Underwood & Schulz, 1960).

In this paper, I group information processing into three stages: *stimulus identification* (stimulus encoding and categorization), response selection, and *response execution* (programming: loading, planning, parameterizing; executing). This grouping forms the basis for my framework of studying practice schedule effects on skill acquisition.

1.4 Framework for Investigating Practice Schedule Effects in Skill Acquisition

Based on a review of the literature, a theory of practice schedule effects on components of task performance is proposed. After outlining the theory, I review evidence in support of the two main points. Practice schedule affects skill acquisition primarily by affecting the stimulus identification and response programming stages. A blocked practice schedule affects stimulus encoding and categorization, improving immediate performance on concept learning tasks (e.g., Elio & Anderson, 1981; Kurtz & Hovland, 1956). However, after a delay a random practice schedule might show better categorization. Practice schedule affects response-oriented component of skill acquisition: a blocked schedule reduces response programming demands during acquisition, which in turn reduces learning as measured by retention and transfer (e.g., Carlson & Yaure, 1990; Lee & Magill, 1983)[2].

I argue this theory can provide a framework for uncovering consistent, reliable effects of practice schedule on skilled performance, as long as the components of task performance are isolated to focus on the individual stages of information processing. The stimulus and response level stages, and the empirical evidence supporting their relevance in practice schedule manipulations, are discussed in more detail in the following sections, as they provide the rationale for Experiments 1 and 2 (response level stage) and Experiments 3 and 4 (stimulus level stage).

[2] Practice schedule affects stimulus-response association indirectly. A blocked schedule, due to its massed presentation (Richland, Finley, & Bjork, 2004), impairs S-R association relative to a random schedule, with distributed practice (i.e., a spacing effect; e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Melton, 1967). Separating spacing effects from contextual interference effects is not the focus of this dissertation (but see Richland et al. for one attempt).

1.5 Practice Schedule Effects in Response Execution

Although there is a dearth of research focusing specifically on the stimulus-oriented stages (stimulus identification and categorization) in skill acquisition, many studies have focused on the response-oriented stages (response execution). Especially in motor learning, many of these studies have shown strong practice schedule effects. These occur with a variety of motor tasks, and occasionally cognitive tasks. These experimental tasks, although varied in goals and processing requirements, consistently minimize the stimulus categorization and stimulus-response association demands, instead emphasizing executing the response as the key determinant of proficient task performance.

1.5.1 Task characteristics

Multiple studies using motor (and occasionally cognitive) tasks report results showing practice schedule can have consistent effects on acquisition and retention performance: participants in a blocked practice schedule outperform the random schedule participants during acquisition, however, this pattern reverses itself on retention or transfer trials. This general pattern has been shown using multi-segment movement tasks emphasizing speed (Immink & Wright, 1998; Lee & Magill, 1983; J. B. Shea & Morgan, 1979), multi-segment movement tasks emphasizing relative timing accuracy (Lee & Magill, 1983, Experiment 1; Lee et al., 1992; Li & Wright, 2000; Simon & Bjork, 2001, 2002), open-loop motor tasks (force production tasks; C. H. Shea, Kohl, & Indermill, 1990), and cognitive procedural skills (Carlson & Shin, 1996; Carlson & Yaure, 1990). These experimental tasks and results share in common a central characteristic: the primary determinant of performance (and therefore the primary dependent measures) is execution of the response.

First, the tasks minimize stimulus categorization requirements. The stimuli are highly discriminable, and do not require categorization. Participants do not see example instances that then have to be categorized in order to determine the appropriate response. Instead participants see the category prototype directly. For instance, participants see a diagram of the movement pattern to be performed.

Second, the stimulus and response sets, and thus the S-R mappings to be learned, are often small. For instance, participants might learn three movement patterns (J. B. Shea & Morgan, 1979), four Boolean logic gates (Carlson & Yaure, 1990), or five target forces to produce (C. H. Shea et al., 1990).

Third, these S-R mappings are usually present for the entire set of acquisition trials (and thus the S-R is likely overlearned by the end of acquisition). As a result, the participant's primary means of optimizing performance is through efficient and error-free execution of the response.

Finally, the dependent measures emphasize proficient response production, rather than response accuracy. By the end of acquisition both practice schedule conditions generally perform at a high accuracy level; what differentiates the groups is the speed or timing accuracy with which they execute the response (e.g., the speed with which one can knock down a set of three barriers in the prescribed order). Depending on the task design and dependent measures collected, practice schedule affects reaction time (e.g., Lee & Magill, 1983), movement time (e.g., J. B. Shea & Morgan, 1979), and response latency (e.g., Carlson & Yaure, 1990) in speeded tasks, and movement timing accuracy in relative timing tasks (e.g., Lee et al., 1992).

1.5.2 Locus of Practice Schedule Effects: Response Execution

This set of results showing differences in reaction time, movement time, and relative timing accuracy suggest that practice schedule can affect response production. Reaction time is hypothesized to reflect, among other processes, response programming (e.g., response loading or response planning; Schmidt & Lee, 2005). The effect of practice schedule on movement time was originally hypothesized to arise from participants making and correcting errors during execution of the movements (J. B. Shea & Morgan, 1979). However, further evidence shows that movement time is also affected by ongoing (online) response programming, especially if the task requires executing a sequence of responses (Klapp & Wyatt, 1976), or if participants are not given enough time to study the movement before initiating the response (Immink & Wright, 1998). This online response programming also provides a locus of differences in accuracy in relative timing tasks as a function of practice schedule. If the random schedule increases planning demands then this planning might be occurring during execution, increasing timing errors (Li & Wright, 2000).

1.5.3 Cognitive Processes Affected by Practice Schedule

Several cognitive processes might account for the effect of practice schedule on response programming, including higher demands on attention and working memory (WM). These demands might arise from the need to load and plan a new response pattern on sequential trials. The hypothesis that the locus of practice schedule effects on the response execution involves programing the response (i.e., loading and preparing responses) is supported by a variety of empirical data.

First, the benefit of a random practice schedule results from practicing the tasks in varying sequences from trial to trial, not because of uncertainty about which task will be practiced from trial to trial (Lee & Magill, 1983, Experiments 2 and 3). Using a *serial order*, in which tasks are ordered in a predictable manner such that participants always know which response is required on trial$_{n+1}$, Lee and Magill (1983) showed the contextual interference effect could be produced with either a random or serial practice schedule.

Second, Li and Wright (2000) used a dual-task methodology to show that a random schedule recruits additional attentional resources when planning the response for a different task (i.e., planning for Task A on trial$_{n+1}$, when Task B was just executed on trial$_n$), and that this same attentional demand occurs in a blocked schedule on trials that switch tasks (e.g., between blocks).

Third, the CI effect can occur in a blocked schedule if one adds intervening (non-interfering) tasks that utilize WM processing (e.g., math and same-different judgments) between trials. Acquisition training using intervening, non-interfering tasks reduces response latency on transfer trials that require a WM load (i.e., search plus WM storage). The authors' explanation that the benefit results from an improved "ability to coordinate representations and procedures in WM" (Carlson & Yaure, 1990, p. 494) is consistent with a response programming account of CI effects[3].

Fourth, when trained on single Boolean operator problems, a random practice schedule reduces response latency on transfer problems that require applying a series of *different* Boolean operators (mixed rule problems). This effect of practice schedule does not appear on transfer problems that require applying a sequence of the *same* Boolean operators (i.e., "single rule"). These data suggest random practice improves the fluency with which one can apply different operators (response programs), perhaps arising from differences in the ability to load and manipulate operators in WM and to sequence the application of those operators (Carlson & Shin, 1996).

Finally, in support of the idea that the differences between blocked and random practice schedules occur because of requirements to prepare and execute a response program, when the set of tasks being learned have the same response pattern and require only different parameterizations, practice schedule does not show an effect on acquisition or retention performance (Lee et al., 1992; Magill & Hall, 1990).

[3]In this example, contextual interference arises from the contextual variety in which the task is practiced. However, this variety is operationalized by interleaving non-similar tasks from a different domain, rather than similar tasks from the same domain (as is done with a random practice schedule). This is consistent with the original conception of contextual interference (Battig, 1979)

1.5.4 Demonstrating Practice Schedule Effects in Response Execution

Based on the results showing practice schedule effects on response execution, a series of recommendations for finding practice schedule effects in motor or cognitive tasks that have a heavily weighted response execution component is proposed. First, to clearly see the specified locus of the effect, the stimulus categorization and response selection components of the task should be minimized, while the response component should be emphasized. Second, given the importance of response programming (including both loading and preparing the response), the set of tasks to be trained should use different sequences of responses (such as multiple movements, timings, or application of operators). Thus, cognitive tasks might use different schemas and motor tasks might use different motor programs. Third, the strongest effects of practice schedule are hypothesized to appear in response latency and timing accuracy, not response accuracy. These differences in response latency and timing accuracy arise from differences in response planning and execution time, and are hypothesized to arise from differences in the speed with which on can load and parameterize a response sequence. These three recommendations guide the design and selection of tasks in Experiments 1 and 2.

1.6 Practice Schedule Effects in Stimulus Identification

The majority of practice schedule research in skill acquisition has not focused specifically on stimulus learning (excepting a few studies; Gane & Catrambone, 2009; Quilici & Mayer, 1996, Experiment 3). However, related research on inductive learning and concept learning has occasionally investigated sequence and order effects. A review of these studies provides the theoretical justification for predicting the effects of practice schedule in acquiring skill in stimulus identification.

1.6.1 Blocked Presentation is Beneficial

Several studies suggest that massing example instances according to their category (i.e., a blocked schedule) can improve acquisition of concepts and conceptual rules (Elio & Anderson, 1981; Hiew, 1977; Kurtz & Hovland, 1956). Hiew (1977) taught participants to categorize simple geometric stimuli based upon identifying the bidimensional rule that governed stimulus classification (e.g., "All patterns which are red and square are examples." Bourne, 1970, p. 547). During acquisition, the "systematic" condition (i.e., massed by rule; blocked practice schedule), had lower trials to criterion and lower latency, relative to the "mixed" condition (i.e., rule changed after every problem; random practice schedule). However, the systematic group's performance suffered each time the rule switched.

Research with similar stimuli (geometric figures) shows that massing presentation can increase participants' ability to verbally describe the rules governing classification of instances to categories (Kurtz & Hovland, 1956). Thus, massing not only improves performance, but can make the rules consciously available. Additionally, the rule learning task has been extended by replacing geometric patterns with verbal descriptions, and increasing the difficulty of the classification task by using more ill-defined categories. For example, participants learn to categorize which "club" a person belongs to, following descriptions of other members of the club: "One member of the Dolphin Club is a Baptist, plays golf, works for the government, is college educated, and is single" (Elio & Anderson, 1981, p. 401). Blocking example instances according to the shared features necessary for categorization reduced trials to criterion and increased accuracy of identifying novel items.

This research on sequence effects in concept learning suggest several mechanisms by which order effects might arise. One proposed mechanism is that blocking improves schema abstractions (i.e., generalizations or concept learning) because blocked sequences contain fewer intervening items with different features. Thus, shared features among multiple items are simultaneously present in working memory, which is hypothesized to improve generalizing from instances (Elio & Anderson, 1981; Underwood, 1952).

An alternate mechanism is that the order of example presentation can affect the saliency of the features in successive examples. When successive examples show small variations from trial to trial, individuals have a greater ability to discriminate between new versus old examples, and individuals focus on those dimensions that remain constant across examples (Medin & Bettger, 1994). Additionally, when category membership is determined by the cooccurance of two dimensions, blocking instances by those correlated dimensions increases the tendency for participants to use those correlated dimensions to classify novel instances (Wattenmaker, 1993, Experiment 2). Thus a blocked order can increase the salience of the critical features and increase participants' use of these features, thus speeding category learning.

In addition, the benefit of blocking does not seem to be predicated upon the individual actively searching for commonalities among examples that determine category membership. Research using a variety of task instructions demonstrates the benefit of blocking sequences in both intentional and incidental learning paradigms. Explicitly instructing participants to learn categories or rules for categorization show benefits of blocking (Hiew, 1977; Kurtz & Hovland, 1956; Wattenmaker, 1993). Blocking also improves categorization when participants are dissuaded from hypothesis testing via explicit instructions to memorize instances rather than test hypotheses. Likewise, blocking improves categorization when participants are instructed to rate stimuli for likability, to construct visual images of stimuli, or to memorize example-label associations (Wattenmaker, 1993). Even in these incidental learning conditions, blocking example instances affects the encoding of examples, and thus reveals itself when individuals classify new instances by using analogical mapping to stored examples, rather than explicit rule identification (Wattenmaker, 1993, Experiment 4).

In summary, it seems multiple mechanisms are responsible for the benefits of blocking sequences of presentations, and that these multiple mechanisms allow a blocked schedule to affect performance regardless of whether participants are actively attempting to learn classification rules or processing individual examples (without explicit hypothesis-testing).

1.6.2 Random Presentation is Beneficial

Although the previous studies present strong reasons why blocking might improve concept learning, one recent study has provided some evidence that blocking hinders concept learning (Kornell & Bjork, 2008). Kornell and Bjork investigated perceptual classification (classifying a painting as an instance of a specific painter). Sequentially presenting multiple paintings from the same artist reduced identification accuracy when participants later classified novel paintings. Kornell and Bjork, however, do not present a mechanism for why massing hinders initial concept learning.

1.6.3 Using Skill Acquisition Theories to Inform Concept Learning Results

The research on massing example instances in concept learning paradigms provides compelling evidence that blocking should facilitate immediate concept learning and categorization performance. However, performance post-training is not necessarily indicative of *learning* (Schmidt & Bjork, 1992). When retention and transfer measures are considered, the evidence favoring blocked presentations is lacking. Indeed, it might be that the majority of the concept learning studies overlook critical tests of learning, retention and transfer performance, by equating good initial performance with the "optimal" training paradigm (Battig, 1972; Bjork, 1994; Schmidt & Bjork, 1992; W. Schneider, 1985).

None of the concept learning research cited earlier measured retention. Two studies (Elio & Anderson, 1981; Hiew, 1977) investigated transfer, but only Hiew investigated transfer to a different rule set. The Hiew data showed that although the blocked condition resulted in fewer trials to criterion, the random condition actually performed better on novel bidimensional rules (i.e., complementary to the rule structure learned earlier; near transfer), and on novel tridimensional rules (i.e., a modification to the rules learned earlier; far transfer). In effect, these transfer tasks use strategic knowledge, rather than stimulus-specific knowledge (as Elio and Anderson use).

Evidence from visual discrimination research provides further support for the idea that initial difficulty might improve transfer by demonstrating that the learning examples seen early in training can strongly shape the strategy learners use to make future discriminations. The strategy learners adopt to discriminate between visual stimuli can involve identifying which subset of features of the stimuli are attended. These strategy differences manifest in response latency and accuracy measures of same-different judgments, indicating that strategy affects encoding or comparison processes. Furthermore, this initially learned strategy can persist in later trials and novel tasks (Doane, Alderton, Sohn, & Pellegrino, 1996).

Although Doane et al. (1996) did not look at practice schedule, their results are informative. Suppose that example order makes some features more salient than others (e.g., Medin & Bettger, 1994). These early training example instances have strong effects on the strategies participants use to discriminate among stimuli (e.g., picking which features to focus on). Furthermore, this strategy can persist, creating long-term consequences for their ability to transfer knowledge (Doane et al., 1996). For instance, if the training examples are blocked in a way to key participants to focus on Dimension 1 and Dimension 3, then after successive trials they will tend to use the values of Dimensions 1 and 3 as their classification rule, even in the presence of new stimuli that require different dimensions to be classified (similar to an Einstellung effect).

1.6.4 Demonstrating Practice Schedule Effects in Stimulus Identification

The concept learning research suggests that massing improves acquisition performance. Thus, a blocked practice schedule should result in higher categorization accuracy than a random practice schedule. However, the concept learning literature does not provide strong evidence as to whether this improved acquisition performance extends to retention or transfer performance. Given that a random practice schedule should create interference, one prediction is that if retention and transfer measures are included, a contextual interference effect might appear. Specifically, one should see a pattern such that categorization accuracy is initially better for the blocked schedule group, but that a reversal occurs when retention or transfer items are introduced, and the random schedule group outperforms the blocked schedule group. In order to test these predictions and focus the locus of the effects of practice schedule on the stimulus identification stage, the response selection and response programming stages of the task should be minimized. This framework is used in Experiments 3 and 4 to test practice schedule effects on category learning (conceptual and perceptual learning).

1.7 *Practice Schedule Effects with Varying Amounts of Practice*

Although these explanations of how practice schedule affects processing are promising, they characterize only a single slice of time in the entire skill learning process. As learners progress through stages of skill acquisition their focus of learning changes (Fitts & Posner, 1967); practice schedules might have different effects at different stages. One way to test this prediction is to vary learners' amount of practice, by exposing learners to different amounts of acquisition trials.

I argue that amount of practice is an under-investigated variable in the practice schedule literature and has the potential to:

1. Function as a critical variable that bounds the conditions under which practice schedule effects are observed, thereby supplementing extant theories that CI is the mechanism by which practice schedules affect skill acquisition.

2. Inform decisions in applied situations. That is, when designing instruction for training and education, and given constraints about the amount or duration of training, what practice schedule is optimal?

3. Provide a theoretical grounding for two existing hypotheses about the bounding conditions for contextual interference: complex skills do not show contextual interference effects and domain experience creates attribute treatment interactions.

1.7.1 Theoretical Explanations for Why Practice Schedule Interacts with Amount of Practice

Based on the theories proposed to explain contextual interference effects and the effects of practice amounts on skill acquisition (in general) and skill acquisition (under different practice schedules), I predict an interaction between practice schedule and amount of practice. With *low* amounts of practice, a blocked schedule is either (1) equivalent to or (2) better than a random schedule, in terms of performance after a delay. With *medium* amounts of practice and *high* amounts of practice the blocked schedule is worse than the random schedule, in terms of performance after a delay. Additionally, with high amounts of practice, this difference between the two schedules is greater than the difference between schedules when medium amounts of practice are used. The theoretical rationale for each of these predictions is explained below, which is then followed by a review of the extant data relevant to these predictions.

1.7.1.1 Low Practice

In low amounts of practice, two different predictions have been proposed: (1) a blocked schedule is equivalent to a random schedule, and (2) a blocked schedule is better than a random schedule.

Blocked equivalent to random. The hypothesis that with low amounts of practice, acquisition performance under a blocked schedule will be equivalent to performance under a random schedule is informed by theories of the stages of skill acquisition. When learners are in an early stage of learning they are concentrating on learning the basics of the task demands (e.g., the *cognitive stage*; Fitts & Posner, 1967, or the *encoding stage*; Neves & Anderson, 1981). At this point individuals are learning what to do to perform the task at a minimum level of proficiency (i.e., processing task demands, elaborating on instructions or limited information in the environment, hypothesis testing about what response to execute or how to sequence responses, etc.). In both blocked and random schedules participants are engaging in some form of elaboration and processing of the tasks; this elaboration and processing improves the probability of retrieval in later trials (Battig, 1979; J. B. Shea & Morgan, 1979). In this early stage, manipulating practice schedule either (A) does not dramatically change the type of elaboration or processing, or (B) does change the type of elaboration or processing, but at this stage a range of elaborations and processing are equally beneficial. According to this second argument, at this early stage of learning there is much to be learned, there are many potential retrieval cues, and so virtually any type of task-specific elaborations or processing is beneficial; it does not yet matter whether participants are making inter-example or intra-example comparisons (Quilici & Mayer, 1996; C. H. Shea et al., 1990). Regardless of whether the processing is the same (Theory A) or different (Theory B), both explanations holds that in retention, both conditions (blocked and random) will perform at equivalent levels.

Blocked better than random. A second, alternative hypothesis is that with low amounts of practice acquisition an individual's performance under a blocked schedule will be *better* than performance under a random schedule. In this early stage of learning, individuals are concentrating on learning how to perform the task by focusing on the invariant features of the task. In the blocked condition, because each task is repeated, there are fewer changing features between trials. In the random condition, because tasks change from trial to trial, there are many changing features and thus the learner is overwhelmed and has difficulty focusing on learning to perform the invariant features. Thus, at the end of acquisition the random group is not performing as well as the blocked group *and* has not learned as well as the blocked group (C. H. Shea et al., 1990). This explanation predicts that blocked will be better than random during acquisition (i.e., the standard pattern predicted by traditional explanations of contextual interference effects). Additionally, however, this explanation also predicts that on retention trials the blocked schedule *will still perform better* than the random schedule. This prediction is not made by traditional contextual interference theories (e.g., Carlson & Yaure, 1990; Lee & Magill, 1983; J. B. Shea & Morgan, 1979).

In summary, the first hypothesis argues that early in skill acquisition any practice schedule results in beneficial elaborations and when training ends both groups will perform equivalently. The second hypothesis argues that individuals must learn the basics of the task before a random practice is beneficial. If training ends before the task is learned well enough, then the blocked schedule will perform best on later tests. In this proposal I endorse the second hypothesis: random practice is only beneficial later in practice. Of the studies that manipulate amount of practice and practice schedule, the study with the most clear data and results supported this second hypothesis: blocked is better early in practice.

1.7.1.2 *Medium Practice*

With medium amounts of practice the standard contextual interference hypotheses hold: a blocked schedule is better than a random schedule during acquisition, but this blocked schedule causes worse performance (relative to a random schedule) during retention tests. This result is based on the standard contextual interference effect theories: the elaborative processing view (J. B. Shea & Morgan, 1979) and the reconstruction view (Lee & Magill, 1983) fall in this hypothesis space.

1.7.1.3 *High Practice*

With high amounts of practice the blocked condition is hypothesized to perform better than the random condition during acquisition, but worse than the random condition during retention. Additionally, the difference between schedules in retention increases as practice increases (thus the difference between the schedules is greater than in the medium amount of practice). This is hypothesized to occur because with continued acquisition practice the random group continues to learn to perform the task in a way that will be beneficial when given randomly ordered retention trials. On the other hand, with continued acquisition practice, the blocked group's learning becomes increasingly tied to the blocked schedule, which will be harmful when given randomly ordered retention trials.

This prediction that increasing practice trials does not necessarily improve learning in the blocked group rests on the theory that strategies used early in training persist long after training has stopped (Doane et al., 1996; Stokes, Lai, Holtz, Rigsbee, & Cherrick, 2008), and thus the blocked group becomes increasingly dependent on the blocked presentation and the retrieval cues afforded by the surrounding blocked task items, whereas with further practice the random group continues to use the strategy of multiple and variable processing, use of multiple retrieval cues, and practice preparing the response (C. H. Shea et al., 1990).

1.7.2 Amount of Practice Explains Existing Data

This theory that practice schedule interacts with amount of practice is promising. It makes predictions about amounts of practice, and integrates with theories that the nature of cognitive processing changes with progression through stages of skill acquisition. Furthermore, it provides an overarching theory for two existing hypotheses in the contextual interference literature: (1) complex skills do not show benefits of contextual interference (Wulf & Shea, 2002), and (2) domain experience creates an attribute-treatment interaction (Magill & Hall, 1990; Wulf & Shea, 2002).

1.7.2.1 Complex Skills

Wulf and Shea (2002) have argued that contextual interference effects found in laboratory studies using simple motor skills do not generalize to complex motor skills. Their primary explanation is that in complex skills the task demands are too great and overload the learner, thus eliminating the predicted benefit of a random practice schedule.

Their hypothesis is relevant to the hypothesized practice schedule by amount of practice interaction: early in practice a random practice schedule might overwhelm a learner, even in simple skills. One prediction is that in complex motor skills a greater amount of acquisition practice (e.g., high amount of practice) might be necessary to demonstrate CI effects. That is, for these complex motor skills, random practice schedules might still be useful, but only after a much longer amount of practice than needed with simple skills. This theory that practice schedule interacts with amount of practice then provides an empirical test for the Wulf and Shea hypothesis: rather than varying the complexity of task domain to be trained to show CI effects appear and disappear, one can pick a complex task and vary the duration of the training, which should make CI effects appear and disappear. Only in situations with higher amounts of practice will a random schedule show benefits over a blocked schedule.

One advantage of this approach is that it provides a solution to a dilemma voiced by Wulf and Shea: defining the *complexity* of a task is a difficult (and often circular) proposition, which hinders explicit tests of their hypothesis.

1.7.2.2 Experience as an Attribute-treatment Interaction

Experience in a task domain also seems to moderate the effect of practice schedule (for a review, see Magill & Hall, 1990; Wulf & Shea, 2002). In their review, Wulf and Shea argued that several studies show that with low experience in a domain, a blocked schedule, not a random schedule, improves learning. Additionally, they argued this hypothesis is supported by several studies that show an advantage of a blocked schedule for children but not adults, and argued that this reflects differences in experience. In general, the same argument used for complexity applies to experience: with low experience even simple tasks ordered according to a random schedule might overload the learner. In contrast, with higher levels of experience, the attention demands are lower and so individual might profit from the additional cognitive demands of a random schedule. As Wulf and Shea acknowledged, however, systematic research in this area is lacking. A longitudinal approach, in which participants are trained through various amounts of practice (and thus experience levels) could contribute to the theoretical hypothesis proposed in their review.

In sum, this theory of the interactive effects of practice schedule and amount of practice makes predictions about changes in processing that occur with practice schedules and explains existing hypothesis regarding contextual interference effects moderated by complexity and expertise. In spite of this, few studies have experimentally tested the predictions of the theory. Those that have yield different patterns of results. These empirical results are reviewed here.

1.7.3 Existing Research on Practice Schedule Effects Under Varying Amounts of Practice

My review of the literature found that of the practice schedule research conducted after the initial studies proposed a contextual interference mechanism for practice schedule effects (e.g., Battig, 1979; Lee & Magill, 1983; J. B. Shea & Morgan, 1979), only three have varied the amount of practice while also manipulating practice schedule (Table 1 presents an overview of the studies). In addition to blocked and random schedules, these studies also use blocked-repeated and serial schedules. Table 2 presents a schematic of these four practice schedules and Figure 1 shows a visual representation of the same practice schedules. C. H. Shea et al. (1990) manipulated amount of practice (low, medium, and high) using an isometric dynamic strength task in which participants learned to produce five different forces. Proteau, Blandin, Alain, and Dorion (1994) manipulated amount of practice (low, medium, and high) using a multisegment movement task, in which participants learned to execute three different movement sequences each with a different goal time. Giuffrida, Shea, and Fairbrother (2002) manipulated amount of practice (low and high) using a multisegment timing task in which participants learned to execute a movement sequence according to one of three different goal times with each portion of the movement sequence requiring a different proportion of the total movement goal time. In the following section I synthesize the results of these studies by research questions. Specifically, how does practice schedule interact with amount of practice as measured during acquisition, retention, and transfer? Unfortunately, the answer to these questions is unclear given the varied results that emerge from the three studies.

1.7.3.1 Question 1. In acquisition, at what point are the performance curves of each practice schedule separable and at what point do the curves converge?

An initial step in understanding how amount of practice interacts with practice schedule is to analyze performance curves during acquisition. Absent analysis of retention

Table 1: Overview of the three practice schedule studies that manipulated amount of practice.

Source	Task domain	DVs	Schedules
Giuffrida et al. (2002)	Multisegment timing	ACE; AE(prop)	Blocked; Constant; Serial
Proteau et al. (1994)	Multisegment movement	RMSE	Blocked; Blocked-repeated; Random
C. H. Shea et al. (1990)	Isometric dynamic strength	E	Blocked-repeated; Random

Note. E is equivalent to RMSE (Schmidt & Lee, 2005). A constant schedule presents only one task, rather than multiple tasks.

Table 2: Sample order of three tasks across trial blocks in either a blocked, blocked-repeated, serial, or random practice schedule.

	Practice Schedule			
Block	Blocked	Blocked-repeated	Serial	Random
1	A_{1-12}	A_{1-12}	A_1, B_1, C_1, A_2, B_2, C_2, ..., C_4	A_1, B_1, C_1, B_2, A_2, B_3, ..., A_4
2	A_{13-24}	A_{13-24}	A_5, B_5, C_5, A_6, B_6, C_6, ..., C_8	A_5, C_5, A_6, C_6, B_5, C_7, ..., B_8
3	A_{25-36}	B_{1-12}	A_9, B_9, C_9, A_{10}, B_{10}, C_{10}, ..., C_{12}	B_9, A_9, A_{10}, B_{10}, A_{11}, A_{12}, ..., C_{12}
4	A_{37-48}	B_{13-24}	A_{13}, B_{13}, C_{13}, A_{14}, B_{14}, C_{14}, ..., C_{16}	A_{13}, C_{13}, C_{14}, A_{14}, A_{15}, C_{15}, ..., A_{16}
5	B_{1-12}	C_{1-12}	A_{17}, B_{17}, C_{17}, A_{18}, B_{18}, C_{18}, ..., C_{20}	B_{17}, A_{17}, C_{17}, A_{18}, C_{18}, B_{18}, ..., C_{20}
6	B_{13-24}	C_{13-24}	A_{21}, B_{21}, C_{21}, A_{22}, B_{22}, C_{22}, ..., C_{24}	B_{21}, C_{21}, C_{22}, A_{21}, B_{22}, A_{22}, ..., C_{24}
7	B_{25-36}	A_{25-36}	A_{25}, B_{25}, C_{25}, A_{26}, B_{26}, C_{26}, ..., C_{28}	A_{25}, C_{25}, B_{25}, B_{26}, A_{26}, C_{26}, ..., C_{28}
8	B_{37-48}	A_{37-48}	A_{29}, B_{29}, C_{29}, A_{30}, B_{30}, C_{30}, ..., C_{32}	C_{29}, C_{30}, A_{29}, B_{29}, B_{30}, A_{30}, ..., A_{32}
9	C_{1-12}	B_{25-36}	A_{33}, B_{33}, C_{33}, A_{34}, B_{34}, C_{34}, ..., C_{36}	C_{33}, B_{33}, B_{34}, A_{33}, A_{34}, C_{33}, ..., A_{36}
10	C_{13-24}	B_{37-48}	A_{37}, B_{37}, C_{37}, A_{38}, B_{38}, C_{38}, ..., C_{40}	A_{37}, B_{37}, A_{38}, A_{39}, C_{37}, C_{38}, ..., B_{40}
11	C_{25-36}	C_{25-36}	A_{41}, B_{41}, C_{41}, A_{42}, B_{42}, C_{42}, ..., C_{44}	B_{41}, A_{41}, C_{41}, B_{42}, A_{42}, A_{43}, ..., C_{44}
12	C_{37-48}	C_{37-48}	A_{42}, B_{42}, C_{42}, A_{43}, B_{43}, C_{43}, ..., C_{48}	A_{45}, C_{45}, C_{46}, A_{46}, C_{47}, B_{45}, ..., C_{48}

Note. Each task is denoted by a unique letter. Subscripts indicate trial number for each task.

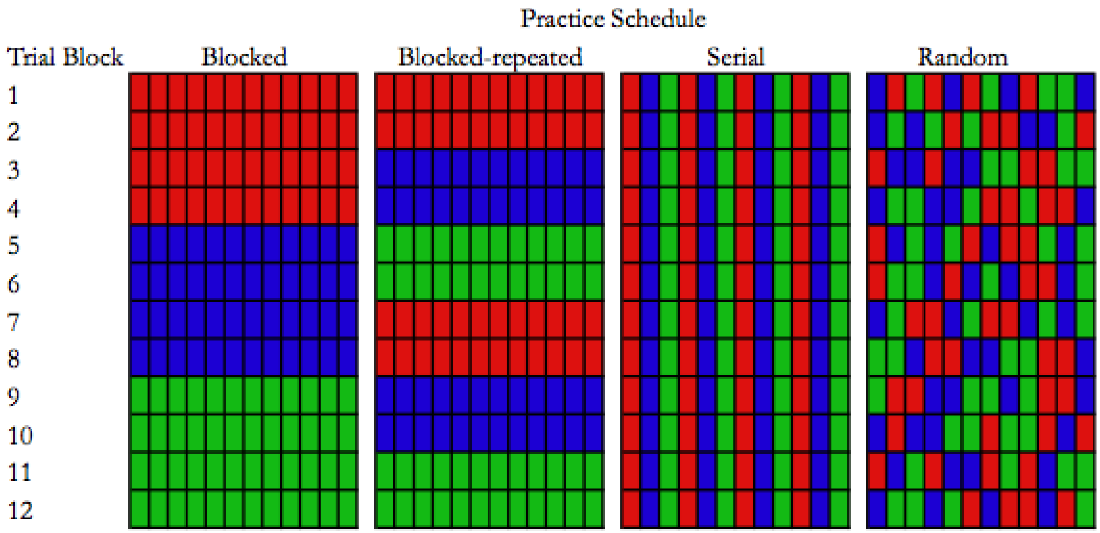

Figure 1: A schematic of the blocked, blocked-repeated, serial, and random practice schedules using three tasks. Each task is denoted by a unique color (Task A = Red; Task B = Blue; Task C = Green). Each cell represents one trial; a trial block contains 12 trials.

or transfer, the acquisition patterns can hint at what effects practice schedule has on immediate performance as learners move through different stages of practice. For example, if in initial stages of learning both schedules use the same processing and elaborations, then early in practice their performance curves should be similar.

Unfortunately, results from the three amount of practice studies with practice schedule have widely varied and so clear patterns do not emerge. In the C. H. Shea et al. (1990) study, the performance curves for movement timing accuracy for the blocked-repeated and random schedules were separate throughout acquisition, with the random schedule always performing worse (see Figure 2). Giuffrida et al. (2002) showed two opposing patterns depending on which dependent measure was used. When using ACE (an error measure based on timing accuracy for executing the entire movement pattern), their results looked like C. H. Shea et al.: by the end of practice ACE seems to have two separable performance curves that do not converge (blocked had lower error than serial; see Figure 3). One caveat, based on their reported statistics, is that it is unclear if the differences early in practice are reliable.

Their other dependent measure, *AE(prop)*, (an error measure based on relative timing accuracy for each segment of the movement pattern), showed a different pattern: throughout acquisition there was never a reliable difference between blocked and serial practice schedules (see Figure 4). The curves appear to start diverging late in practice, however, suggesting that with more practice differences might emerge. Finally, the performance curves reported by Proteau et al. (1994) were separable early in acquisition (with random performing worse), but seemed to converge by the end of acquisition practice (see Figure 5). In sum, the set of studies showed different acquisition patterns across experiments and within experiments (when multiple dependent measures are used).

1.7.3.2 Question 2. When retention trials are ordered using a random schedule, does the amount of practice affect which acquisition schedule causes the best retention performance?

When testing retention, one can order retention trials using a random or blocked schedule. The two studies that have used random order of retention trials reported different results. C. H. Shea et al. (1990) showed an interaction between practice schedule and amount of practice: in low amounts of practice a blocked schedule improved retention performance but in medium and high amounts of practice a random schedule improves retention performance[4] (see Figure 6). On the other hand, the Proteau et al. (1994) data showed a main effect of practice schedule but no interaction with amount of practice: the blocked schedule performed significantly worse than blocked-repeated and random schedules, and the blocked-repeated and random schedules did not differ (see Figure 7).

[4] The C. H. Shea et al. (1990) manipulation confounded practice schedule with amount of practice. In their low amount of practice condition they used a blocked schedule but at medium and high amounts of practice they used a blocked-repeated schedule. Despite this, their random group performed better than the blocked-repeated, suggesting that if the had used a "pure" blocked schedule their observed differences might have been even larger.

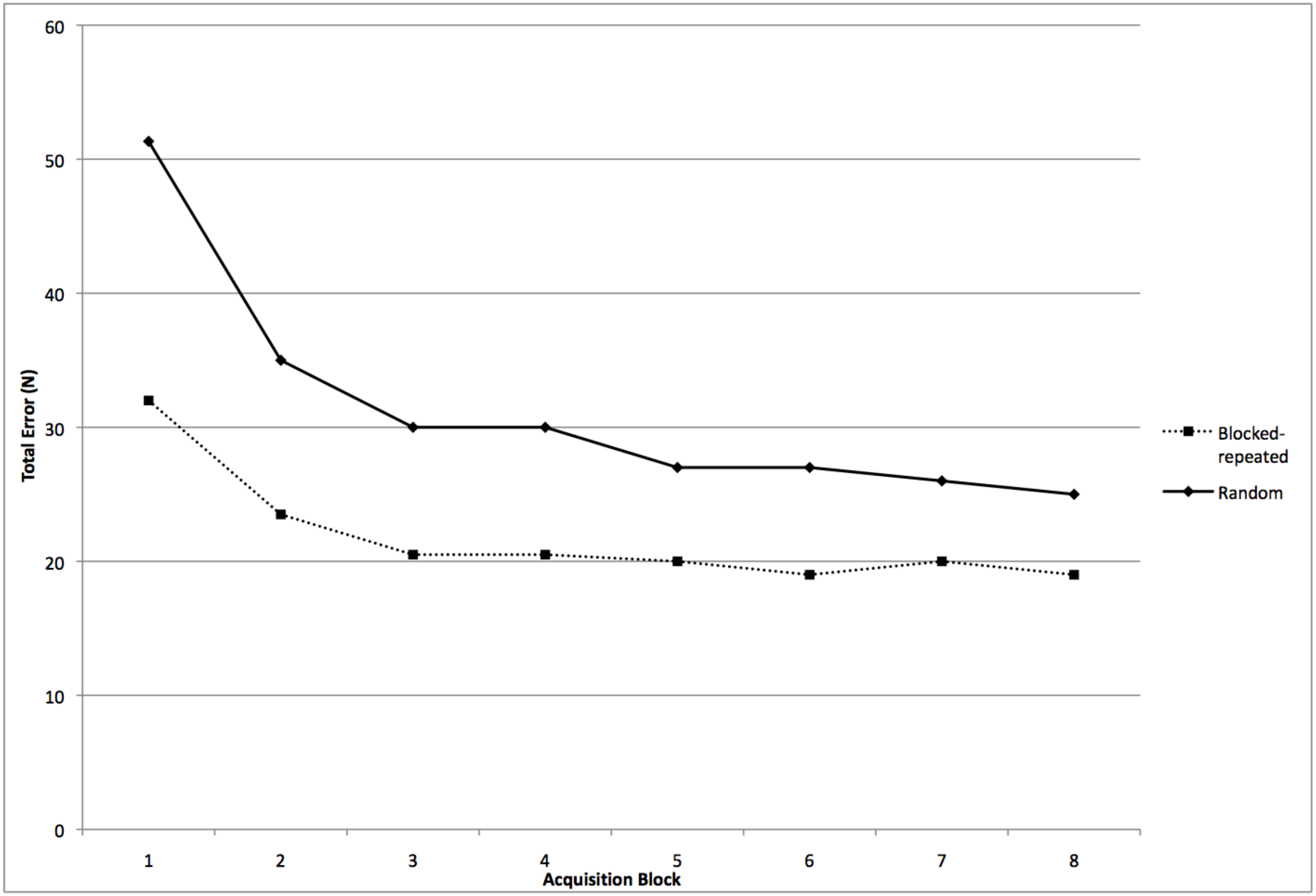

Figure 2: Shea et al. (1990) acquisition results, E.

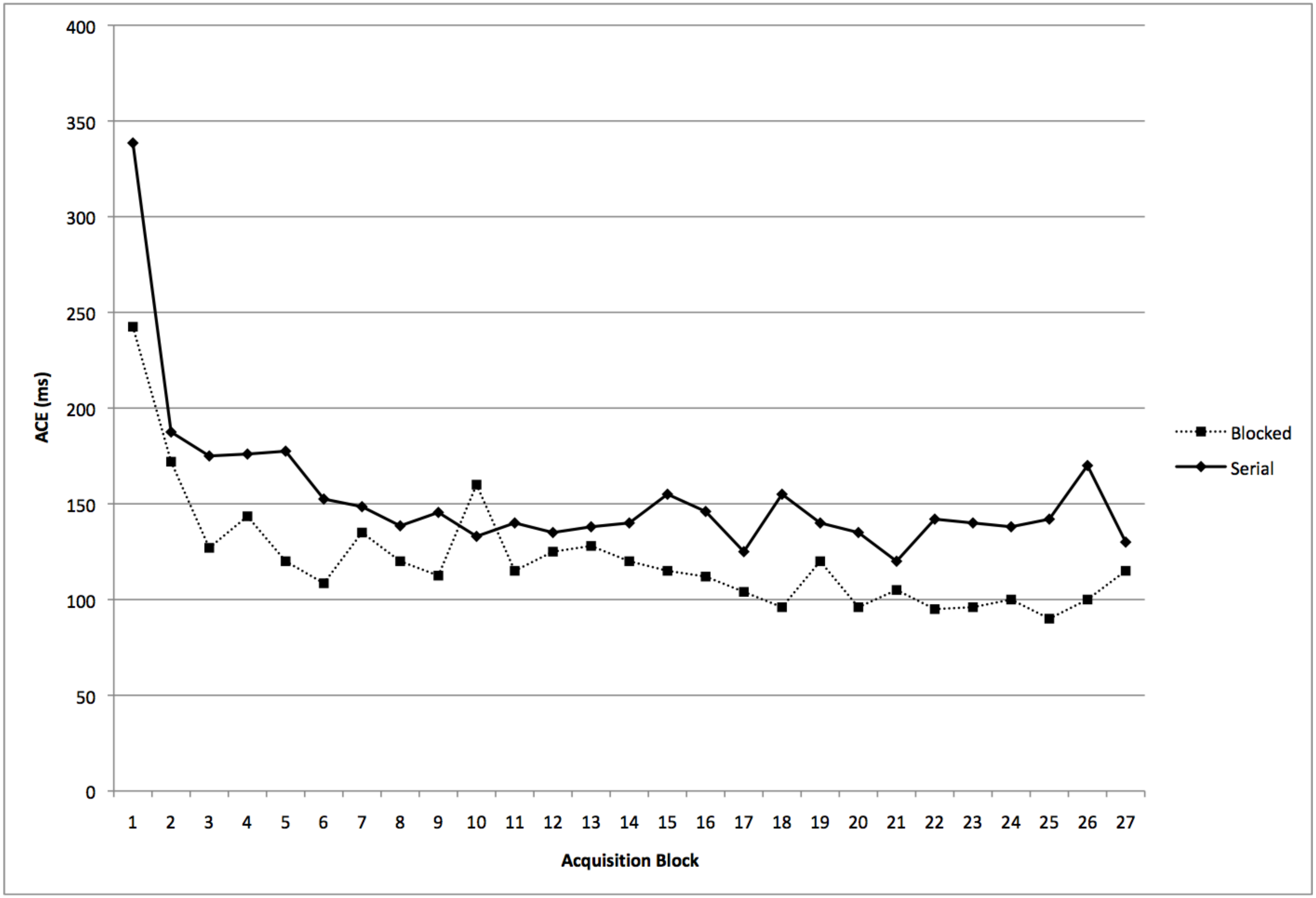

Figure 3: Giuffrida et al. (2002) acquisition results, ACE.

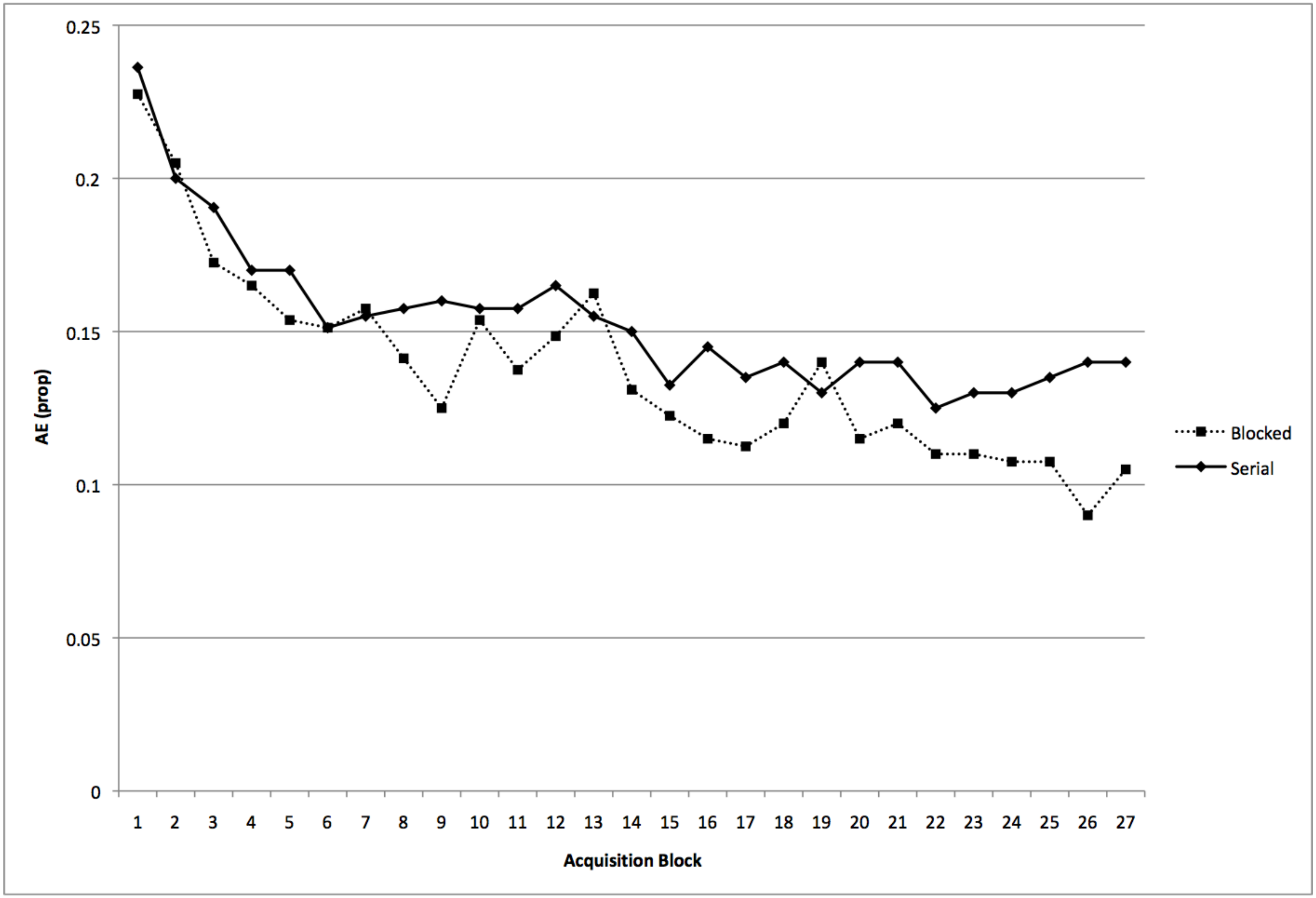

Figure 4: Giuffrida et al. (2002) acquisition results, AE(prop).

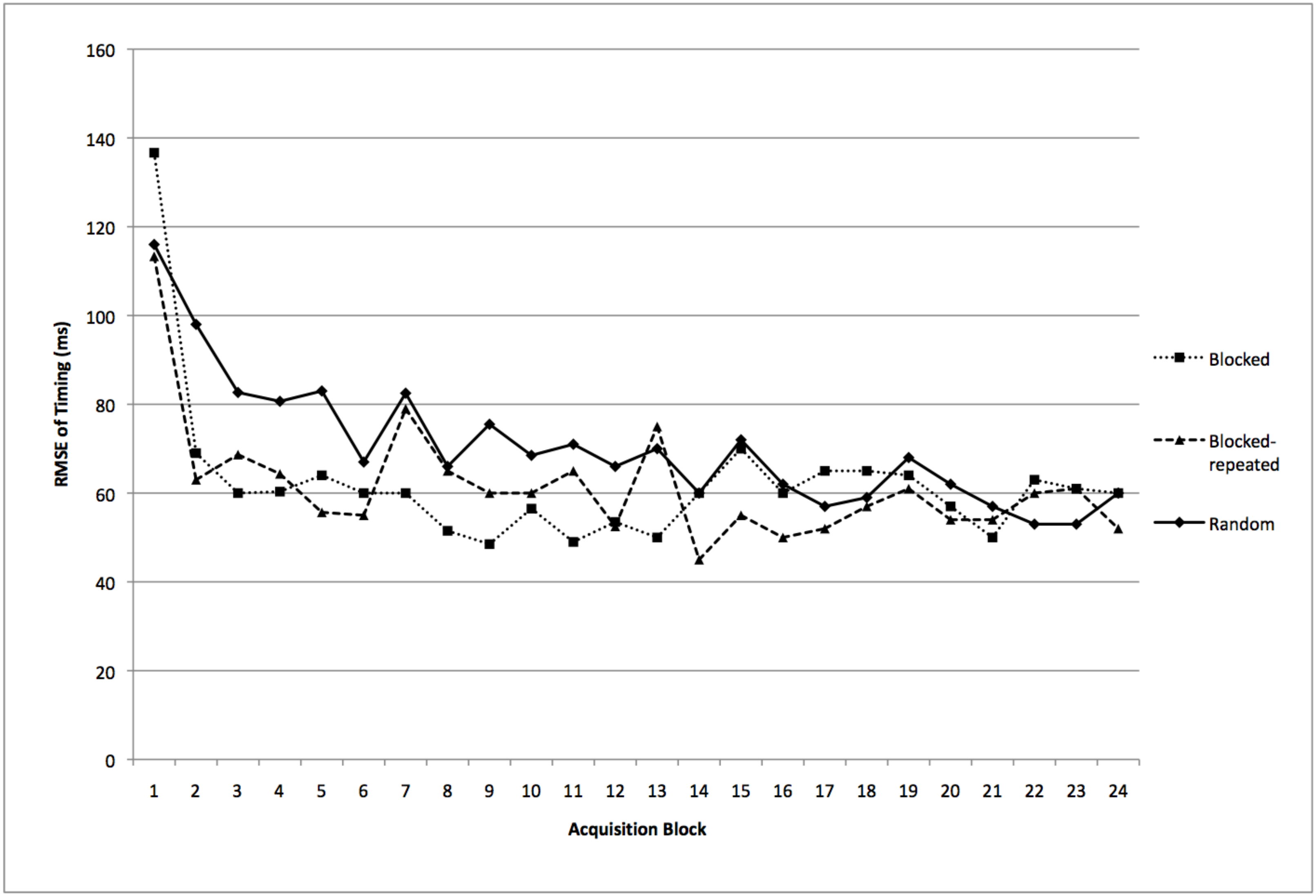

Figure 5: Proteau et al. (1994) acquisition results, RMSE.

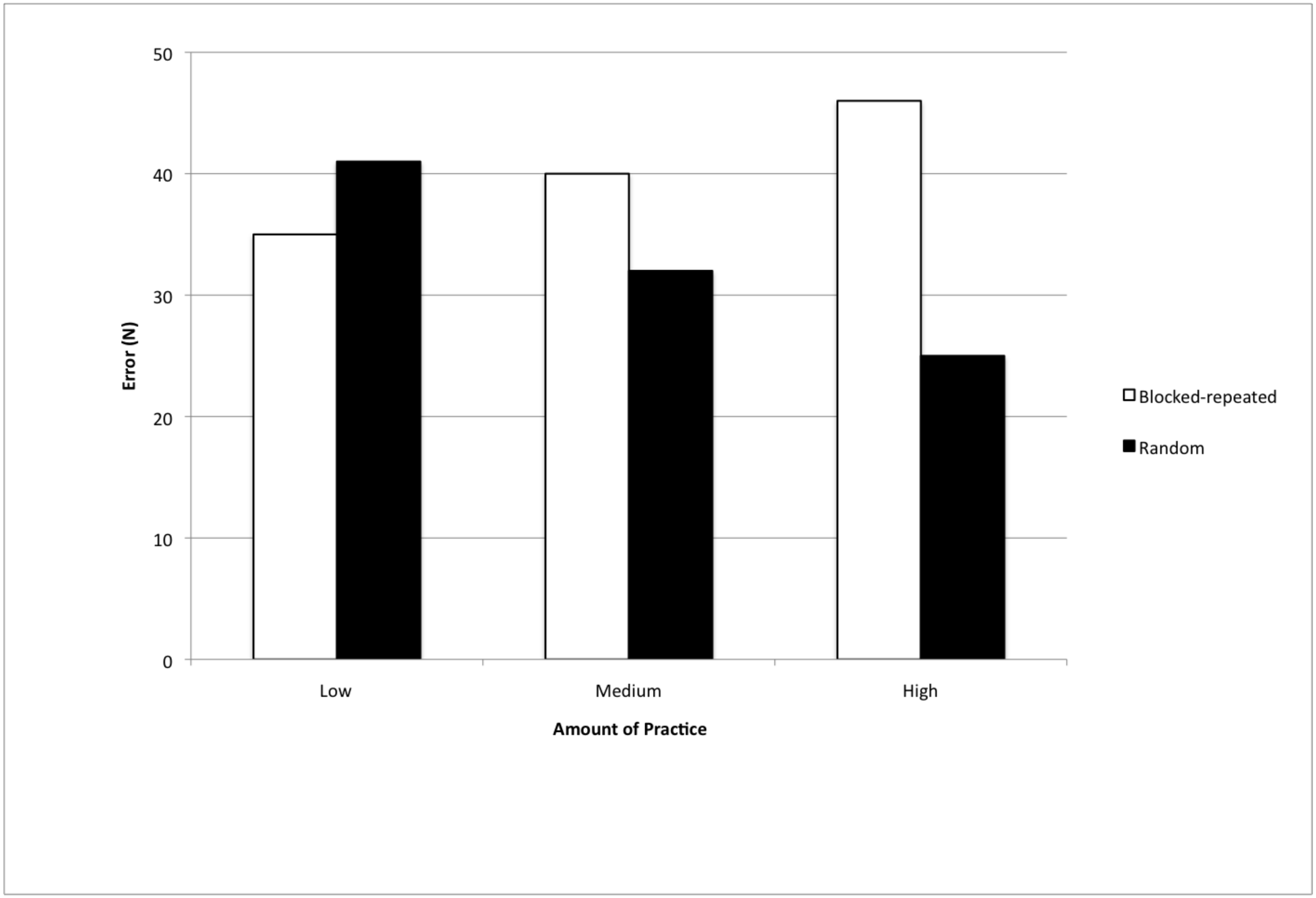

Figure 6: Shea et al. (1990) retention results for retention trials ordered according to a random schedule, E.

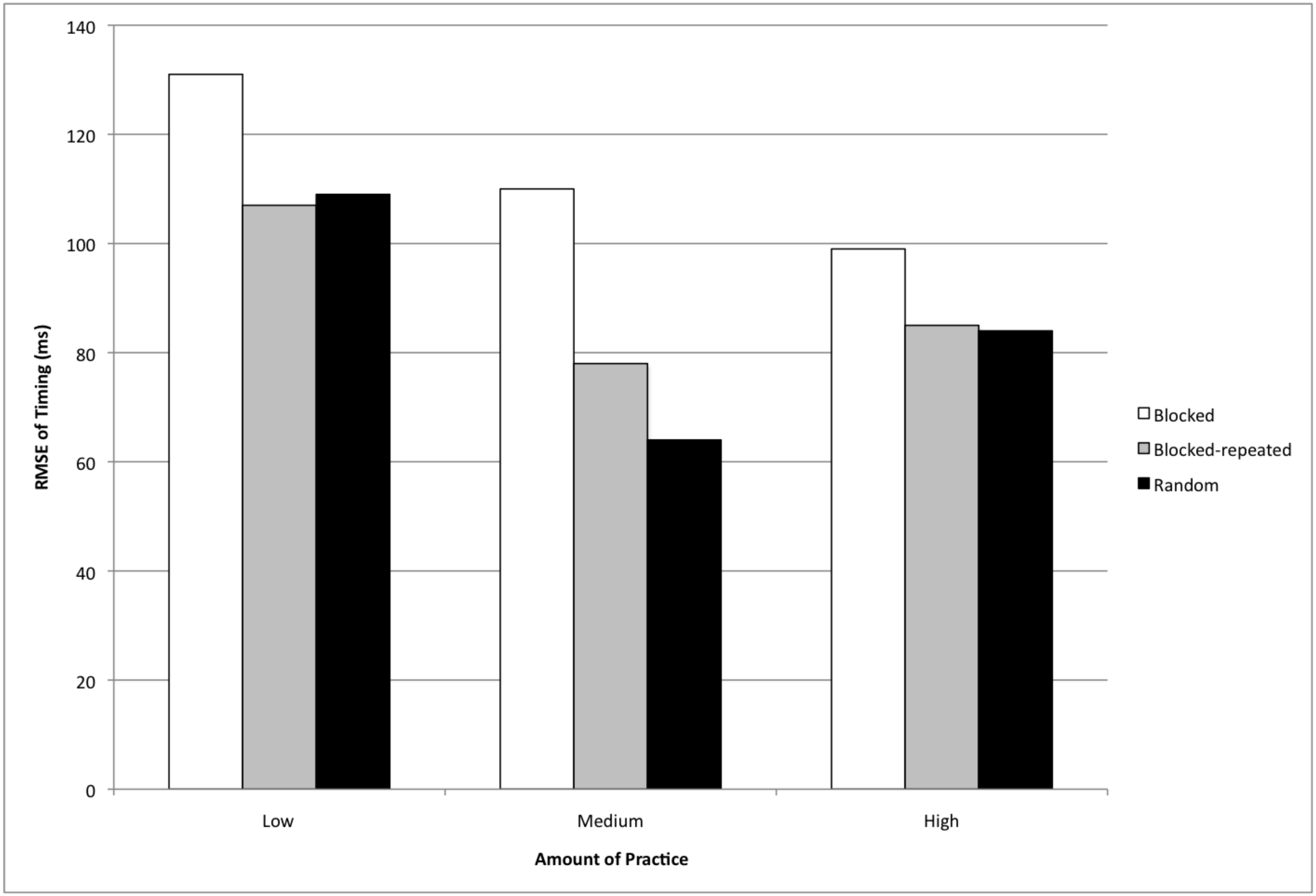

Figure 7: Proteau et al. (1994) retention results for retention trials ordered according to a random schedule, RMSE.

1.7.3.3 *Question 3. When retention trials are ordered using a blocked schedule, does the amount of practice affect which acquisition schedule causes the best retention performance?*

Only the C. H. Shea et al. (1990) study ordered retention trials according to a blocked schedule. With high amounts of practice a random schedule outperformed a blocked schedule during retention, but there was no difference between practice schedules for the low and medium amounts of practice conditions (see Figure 8). Integrating both the random and blocked order retention data from C. H. Shea et al. (1990), random acquisition schedule conditions perform better than the blocked acquisition schedule conditions on retention trials. The amount of acquisition practice needed before the random schedule condition begins to outperform the blocked schedule condition, however, depends on whether retention trials are ordered according to a blocked or random schedule.

1.7.3.4 *Question 4. Does increasing acquisition trials result in increased retention performance?*

The standard answer from the overlearning literature is that increasing acquisition training increases retention performance (Driskell, Willis, & Copper, 1992). An open question is whether practice schedule interacts with this finding. Proteau et al. (1994) found no interaction, only a main effect: increasing training increased retention performance (see Figure 7). C. H. Shea et al. (1990), however, showed that amount of practice interacts with practice schedule. In a blocked-repeated acquisition schedule, increasing practice *increased* error or retention trials ordered according to a random schedule, while in a random acquisition schedule, increasing practice decreased error on those same retention trials (see Figure 6).

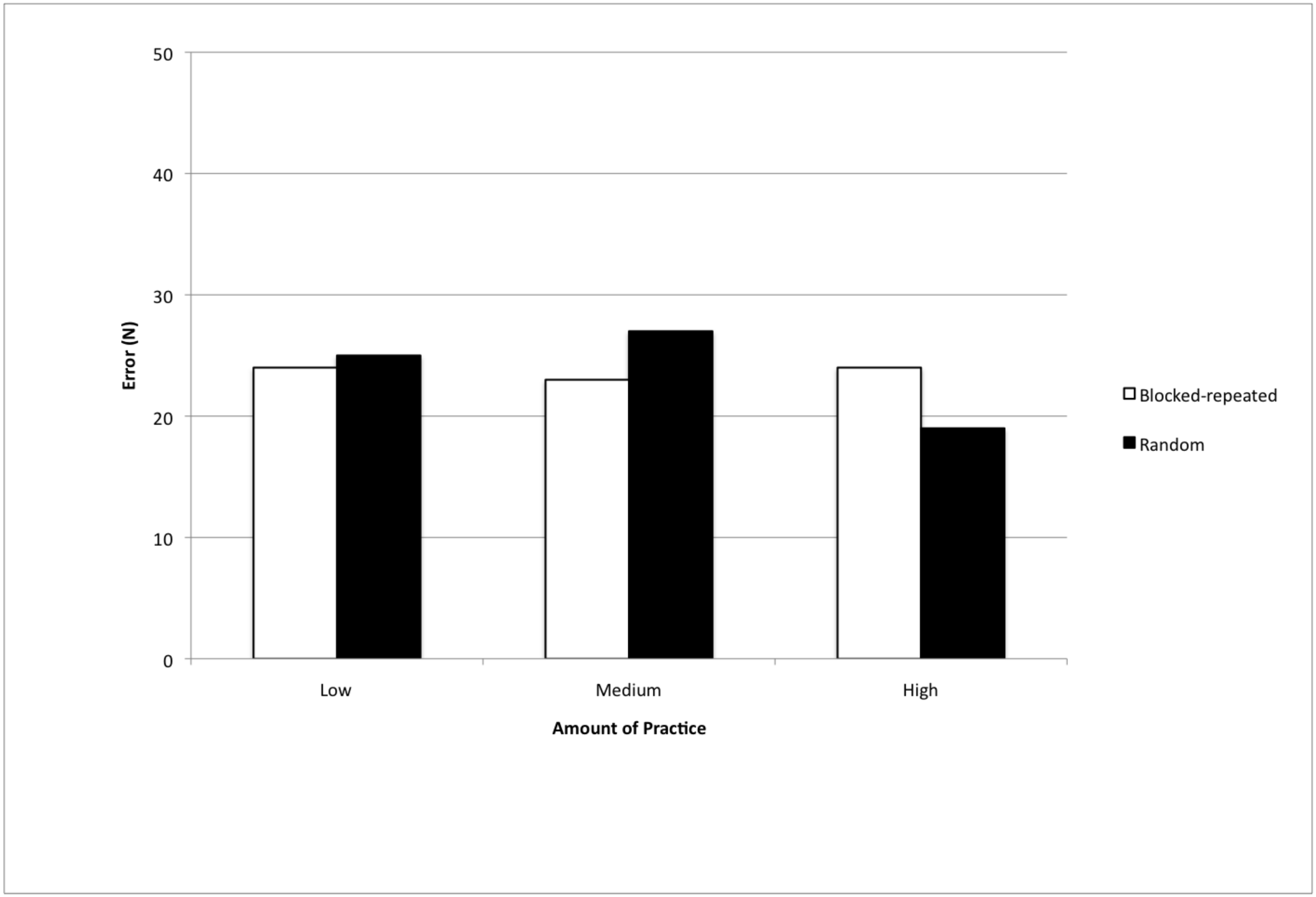

Figure 8: Shea et al. (1990) retention results for retention trials ordered according to a blocked schedule, E.

1.7.3.5 Question 5. When transfer tests are used, does the amount of practice affect which acquisition schedule causes the best performance?

Only one study has manipulated practice schedule and amount of practice and also measured transfer performance. Across both transfer tests (Tests 2 and 3), there was no interaction between practice schedule and amount of practice (Giuffrida et al., 2002).

1.7.3.6 Question 6. Does increasing acquisition trials result in increased transfer performance?

Only Giuffrida et al. (2002) measured transfer performance while manipulating amount of practice. Increasing amount of practice increased transfer performance, but only for one of two dependent measures (AE(prop)) and in only one of two transfer tasks.

1.7.3.7 Summary

Theoretical predictions for how and why practice schedule should interact with amount of practice center on the cognitive processing that occurs during different stages of learning and how that processing might be influenced by practice schedule. Despite these theoretical predictions, the only three studies that have investigated the relationship between practice schedule and amount of practice showed inconsistent results. Most strikingly, all three studies show different acquisition patterns. In general, practice schedules do affect the performance curves, however, the point at which they result in divergent curves differs. Likewise, the retention data also show conflicting results, with one study showing an interaction between schedule and amount of practice (C. H. Shea et al., 1990) and another showing no interaction (Proteau et al., 1994). These inconsistent results may be due in part to differences in designing the acquisition training cutoffs for low vs. medium vs. high amounts of practice, and additionally, between the different tasks used across the three studies. For instance, differences in the complexity of tasks used across studies might result in the patterns

Table 3: The three studies that manipulated practice schedule had different total trials and trials per task.

Source	Total trials	Tasks	Trials per task
Giuffrida et al. (2002)	54; (81); 162	3	18; (27); 54
Proteau et al. (1994)	54; 108; 216	3	18; 36; 72
C. H. Shea et al. (1990)	50; 200; 400	5	10; 40; 80

in the acquisition data, and differences in the number of trials used for low, medium, and high practice might be causing differences in retention performance results.

1.7.4 Critique of Existing Research

The three studies that varied amount of practice are difficult to compare due to differences in (1) defining the amounts of practice, number of tasks, and practice per task, and (2) the nature of the tasks. Although attempts have been made to manipulate amount of practice, this has not been done systematically, making it difficult to find a coherent pattern of results.

1.7.4.1 Different Amounts of Practice Across Studies

Comparing the three studies is difficult because each used different number of total acquisition trials, tasks, and acquisition trials per task (see Table 3). As an example, C. H. Shea et al. (1990) used the largest range of practice amounts (10, 40, and 80 trials per task). Their lowest and highest levels of practice were lower and higher (respectively) than the two other studies. Furthermore, all three studies used different task domains and none reported how the authors derived the number of practice trials that were selected for the groups.

In Table 4 I use the middle amount of practice condition as a baseline to calculate the percentages of practice amount for all studies. This table shows how the percentage of training in the low amount of practice condition varies between studies. C. H. Shea et al. (1990) used 25%, Giuffrida et al. (2002) used 66%, and Proteau

Table 4: The three studies that manipulated practice schedule, with low and high amounts calculated as a percentage of the middle amount of practice.

Amount	Giuffrida et al. (2002)	Proteau et al. (1994)	C. H. Shea et al. (1990)
		Trials per task	
Low	18	18	10
Medium	(27)	36	40
High	54	72	80
		Trials, % of medium	
Low	66%	50%	25%
Medium	(100%)	100%	100%
High	200%	200%	200%

Note. Parenthesis denote an implicit definition because the study did not use a medium amount condition.

et al. (1994) used 50% of the middle condition. C. H. Shea et al. uses the lowest percentage of practice and is also the only study to show a cross-over interaction: with low amounts of practice blocked is better and with high amounts of practice random is better. One empirical question remains unanswered: if Giuffrida et al. and Proteau et al. had used fewer trials in their low amount condition, would they have seen the same results as C. H. Shea et al.?

1.7.4.2 Different Motor Tasks

Comparing the motor tasks used in the three studies is difficult. One method is to compare them on the complexity of the movement patterns that participants are learning. The C. H. Shea et al. (1990) task, a force production task, requires a preprogrammed response; participants hit the arm of a force transducer lever with one hand. This relatively simple movement involves a single response. The Proteau et al. (1994) task requires a responses that may be preprogramed and adjusted during execution. Participants use one hand to knock down wooden barriers located to the left and right of the central axis of the participant, with the constraint that the time between beginning and ending the *total* movement must match the goal movement time. The Giuffrida et al. (2002) task adds an additional layer of complexity to the Proteau et al. (1994) task. In addition to matching the overall goal time of the movement sequence, the time between each component of the movement sequence has a goal time (a proportion of the total sequence's goal time; e.g., Task A: 200–400–300 ms vs. Task B: 250–500–375 ms).

If one uses the above scheme to order complexity of the task domain (e.g., simple to more complex: C. H. Shea et al. (1990); Proteau et al. (1994); Giuffrida et al. (2002)), and compares this to the number of practice trials (see Table 4), one can see that as the studies use more complex movement patterns, the number of acquisition trials in the high condition *decreases*. This results in high amount conditions where participants have less practice with a more complex task. This difference in task difficulties and number of acquisition trials might be one cause of the inconsistent results across the three studies.

1.7.5 Solutions to the Critique of Existing Studies

My dissertation addresses both of the primary critiques of previous studies in this area. First, I use a method for defining amount of practice that can be applied across different domains. Second, this same method of defining amounts of practice and the same general acquisition, retention, and transfer measures will be used in two different task domains: a perceptual categorization task and a motor task.

1.7.5.1 Task-independent Method of Defining Amount of Practice

One contribution of this research is the implementation of a general method for defining amount of practice: a method that can explicitly and systematically define low, medium, and high, and ensures that these amounts retain their meaning when applied in different task domains.

One potential method for defining amount of practice is to set the number of trials for the medium amount based on a criterion measure of performance (e.g., accuracy or response time), and then define low and high as a percentage of this medium amount. For instance, determine the number of trials (on average) needed for accuracy to reach 85%, and then define low and high in relation to this number. Implementing this method in conjunction with a practice schedule manipulation is challenging, however, because one schedule might never reach the criterion. Consider Figure 2. The random schedule condition never reaches an equivalent level of performance as the blocked schedule group (i.e., the performance curves do not converge). How then might one define a criterion level of performance when one of the schedules will never reach the criterion level?

I propose a solution using a criterion *rate of performance improvement*. For each practice schedule, the acquisition data will be fit to a learning curve (e.g., Newell & Rosenbloom, 1981). Computing the first derivative of this learning curve yields the rate of performance improvement (i.e., it describes the velocity of the learning function; Johnson, 1980). Criterion rates of performance improvement (C) can then be defined for each amount of practice (e.g., C_{low} = -0.004, C_{medium} = -0.001, and C_{high} = -0.00025). Appendix A describes how to calculate these values from empirical data. Once these C values are specified, one could use existing performance data to compute the first derivative of the learning function in order to determine the number of trials that are needed to reach each C value, and designing experiments using those values.

Because I do not have learning curves for the motor and categorization tasks, I conducted preliminary studies for each task to generate learning curves and the derivatives of these curves. This allowed me to quantify the amount of practice I used in the subsequent studies.

This method for operationalizing amounts of practice is novel but the outcome (low, medium, and high amount of practice conditions) is consistent with how learning studies conceptualize and define acquisition practice. That is, a researcher might select a number of acquisition trials that will provide a reasonable improvement in performance (i.e., a medium amount of practice). Although further practice would likely result in continual improvement, these additional training trials have diminishing returns (Newell & Rosenbloom, 1981). Likewise, defining the low amount as a period with a high rate of performance improvement corresponds with the hypothetical construct of low practice: at this early stage there is still rapid learning on each practice trial. At this point people are still learning the basics of performing the task; small gains in understanding correspond to large improvements in performance (i.e., a high learning rate). Finally, defining high amount of practice as a lower rate of performance improvement value corresponds to the construct of overlearning: continued practice past the point of a criterion level of performance (or, in this case, the rate of performance improvement). Training continues even though additional training trials only slowly contribute to improvements in performance (i.e., the learning rate has slowed substantially). Despite the slow learning rate, these additional acquisition trials attenuate the drop in performance following a retention interval (Driskell et al., 1992)

In sum, the advantages of this approach are that it provides a quantitative definition of amount of practice that is general: conceptually equivalent amounts can be used even if the experimental task changes. This approach avoids using criterion performance as the measure, which allows it to be applied in experiments that have two or more conditions in which the acquisition performance might not converge. Its operationalization is consistent with theories that advocate stages of practice in skill acquisition: the rate of performance improvement becomes an approximation of moving through stages. Furthermore, by disconnecting the definition of amount from the number of trials used, researchers can systematically *test different definitions* of amount of practice (this idea is expanded in Appendix A).

1.7.5.2 *Different Tasks*

I selected two different task domains, but used the same general method of manipulating practice schedule, amounts of practice, and equivalent measures of learning (e.g., retention and transfer tests). This will make comparisons across task domains easier. A significantly different pattern of results between Experiments 1–2 and Experiments 3–4 would suggest that some characteristic of the task domain might contribute to those differences, rather than methodological details such as definitions of amount of practice, retention interval, and transfer tests. If no differences emerge, then it suggests that the inconsistent results previously found in studies of the amount of practice result from nonequivalent manipulations or measures, not the task or domain used.

1.8 Task Domains

Given my review of the theoretical and empirical evidence for practice schedule effects, I submit that it is possible to demonstrate clear effects of practice schedule on two stages of information processing that enable skilled performance in complex skills: stimulus-oriented stages and response-oriented stages. To demonstrate these effects I used two tasks that emphasize different stages of information processing.

Experiments 1 and 2 tested the effects of practice schedule and amount of practice in a multisegment movement task where the goal is timing accuracy. This multisegment movement task emphasized information processing within response-oriented stages with minimum stimulus-oriented demands.

Experiments 3 and 4 tested the effects of practice schedule and amount of practice in perceptual categorization. These experiments used a task inspired by American football. A complex skill, football has been previously been studied in skill acquisition research (Kirlik, Walker, Fisk, & Nagel, 1996; Nagel, 1993; Walker & Fisk, 1995). The visual categorization task required participants to categorize defensive play diagrams. The task emphasized information processing within the stimulus oriented-stages with minimum response-oriented demands.

CHAPTER II

EXPERIMENT 1

2.1 Introduction

Experiment 1 used a multisegment movement task (e.g., Lee & Magill, 1983; Proteau et al., 1994; Simon & Bjork, 2001, 2002), in which participants learn to perform different movement patterns, each with a unique goal time.

Experiment 1 was conducted to determine (1) the number of practice trials for each practice amount condition in Experiment 2 and (2) which form of a blocked practice schedule (i.e., a blocked or a blocked-repeated) to use in Experiment 2.

2.2 Method

2.2.1 Participants

Forty-nine participants over the age of 17 were recruited from the Georgia Tech participant pool. Twelve of these participants were excluded prior to data analysis for a variety of reasons: did not finish the Session 1 training (N = 1), did not return for Session 2 (N = 2), were falling asleep during Session 1 training (N = 3), ignored instructions on how to perform the task (N = 5), and continuously talked and acted disruptive during training (N = 3).

Excluding these participants resulted in 37 participants; cell sizes are displayed in Table 5. Participants' mean age was 20.1 years ($SD = 1.7$). Of the 37 participants, 23 were male, 13 were female, and 1 did not answer. Participants self-reported their dominant hand: 30 were right-handed, 6 were left-handed, and 1 was ambidextrous.

Table 5: Experiment 1. Number of participants (during Session 1, acquisition) in each acquisition schedule and acquisition schedule version cell.

	Version		
Acquisition schedule	1	2	Total
Blocked	7	6	13
Blocked-repeated	6	6	12
Random	6	6	12

2.2.2 Design

2.2.2.1 Practice schedule

Three practice schedules were used: blocked, blocked-repeated, and random. In the blocked and blocked-repeated schedules, each acquisition block used one movement pattern for all trials. In the random schedule, each acquisition block used all four movement patterns (three trials each). Within those acquisition blocks for the random schedule, the order of the keypress sequences was randomized with the constraint that a movement pattern would not repeat for more than two consecutive trials.

To reduce the possibility that any effects of schedule were caused by the specific order of trials practiced, two *versions* of each of the three types of schedules were created. The blocked-repeated schedules used the same order of tasks as the corresponding version of the blocked schedule, but repeated this order three times. Table 5 shows the cell sizes for each version of the three practice schedules.

2.2.2.2 Amount of practice

Amount of practice (i.e., low, medium, and high) was not manipulated. Instead, participants completed a high number of acquisition trials (n = 432).

2.2.2.3 Retention interval

Participants completed two sessions: Session 1 (acquisition), and 48 hours later, Session 2 (retention and transfer).

Table 6: The four movement patterns (keypress sequence and goal movement time) used during the acquisition and mirror transfer tasks in Experiments 1 and 2.

Phase	Goal time	Keypress sequence
Acquisition	900 ms	9–5–1–2–3
Acquisition	1050 ms	2–1–6–4–8
Acquisition	1200 ms	3–6–5–8–4
Acquisition	1500 ms	4–2–5–8–9
Transfer	900 ms (mirror)	7–5–3–2–1
Transfer	1050 ms (mirror)	2–3–4–6–8
Transfer	1200 ms (mirror)	1–4–5–8–6
Transfer	1500 ms (mirror)	6–2–5–8–7

2.2.3 Materials

2.2.3.1 Task

In this multisegment movement task, participants executed one of four movement patterns: a keypress sequence with a specified goal time. Each movement pattern requires pressing five keys on the numeric keypad in a specified order (see Figure 9). The goal is not to perform the keypress sequence as quickly as possible, but instead to perform the keypresses such that the duration of the keypress sequence matches the goal movement time. Each movement pattern had a unique goal time and keypress sequence (see Table 7).

Participants were given instructions to optimize the accuracy of the movement pattern and timing, as well as to learn to perform the movement pattern in the absence of either the movement time or the diagram.

At the start of each trial, a fixation point was displayed for 2 s, followed by the movement pattern diagram (keypress sequence plus goal movement time). The movement pattern diagram remained visible until a keypress sequence was completed (i.e., after any five keys were pressed).

After each trial participants were given feedback: (1) accuracy of the keypress sequence (correct or incorrect), (2) duration of the keypress sequence, and (3) difference

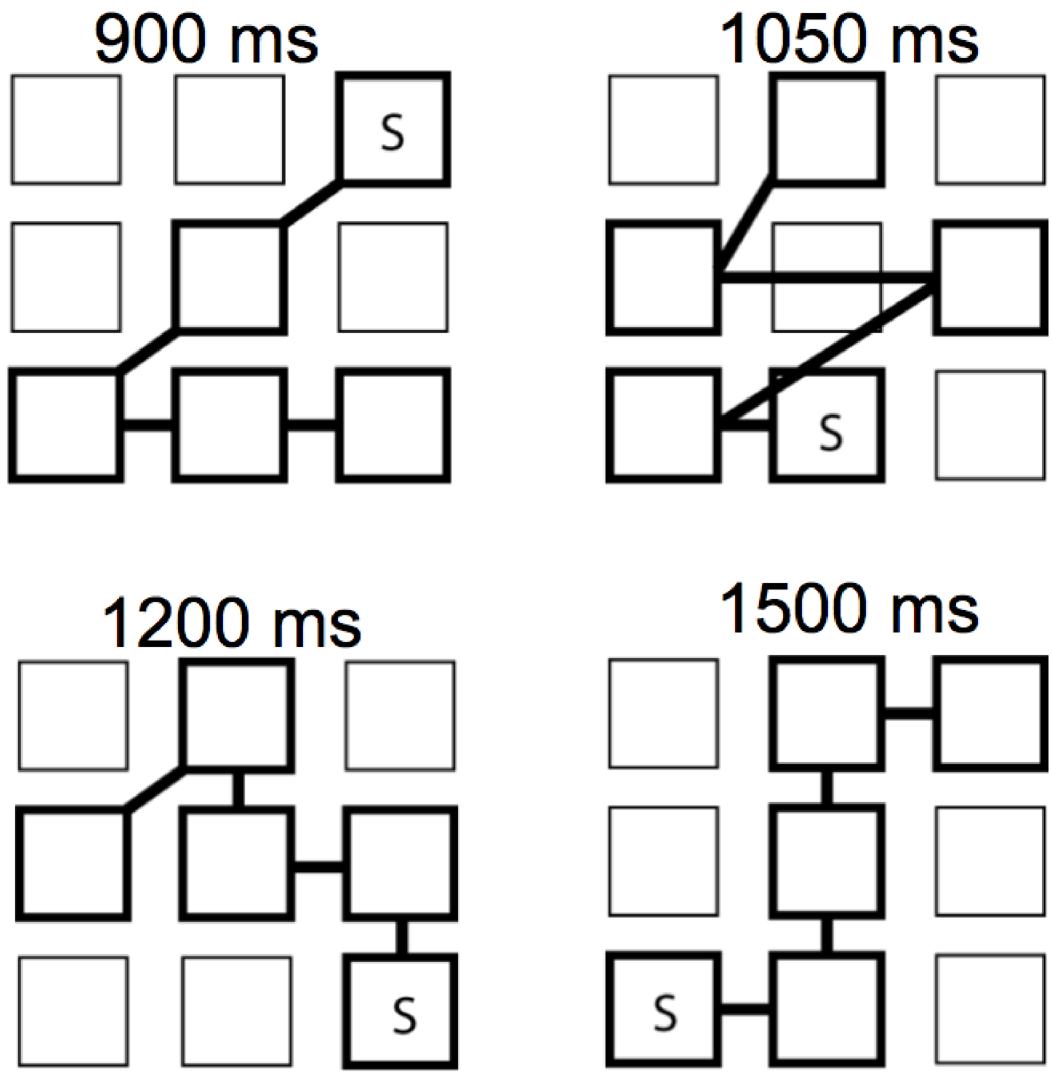

Figure 9: The four movement patterns (keypress sequence and goal movement time) used during acquisition training in Experiments 3 and 4.

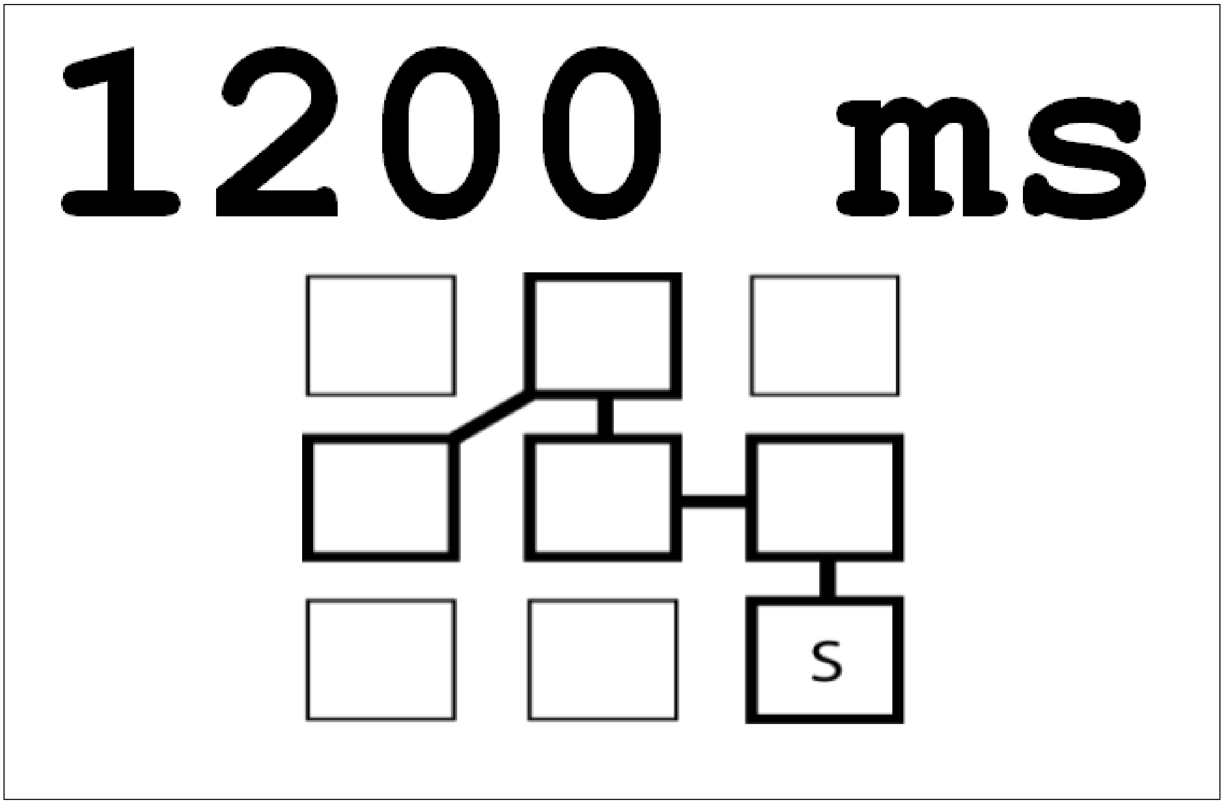

Figure 10: Sample stimuli for the 1200 ms movement pattern in Experiments 1 and 2.

between their movement time and the goal time. Feedback remained on the screen for 5 s[1]. A blank screen appeared between trials (intertrial interval = 2.5 s).

A trial was classified as incorrect if the participant pressed a wrong key in the sequence (e.g., pressed "2" rather than "4"). Incorrect trials were repeated immediately.

2.2.3.2 Stimuli

There were four stimuli screens, one for each movement pattern diagram. Each movement pattern diagram included the relevant keypress sequence and goal time (see Figure 10 for a sample stimuli screen).

[1]Feedback duration decreased with additional acquisition trials: Blocks 1–6 = 5 s, Blocks 7–12 = 4 s, Blocks 13–36 = 3 s.

2.2.4 Procedure

Participants first completed a demographic questionnaire. The experimenter then explained the task, procedure, software, and asked for questions. After all questions were answered the experimenter started the experimental software. The experimental software displayed additional instructions regarding the goal of their task and how to complete it.

2.2.4.1 Session 1

Participants completed 432 correct acquisition trials divided into 3 training epochs with 12 blocks of 12 trials within each epoch (i.e., 36 blocks of 12 trials). Participants received breaks between epochs but, from their perspective, there was no break between blocks. During the break between epochs participants completed a series of paper and pencil tasks (Break 1: the Digit-Symbol Substitution test, Wechsler, 1997, and the Word Beginnings and Endings test, Ekstrom, French, Harman, & Dermen, 1976; Break 2: the Making Sentences test Ekstrom et al., 1976. Breaks were added after piloting to limit fatigue effects; each break lasted approximately 10–12 min. The tasks used during breaks are commonly used as ability measures, but in these studies were used as distractor tasks. The Word Begininings and Endings test and the Making Sentences test were selected primarily because they do not overlap with the cognitive processing required to complete the training tasks.

After completing the three epochs participants completed a recall task in which they had to recall the goal movement time and the keypress sequence for each movement pattern.

2.2.4.2 Session 2

After completing the first session, participants returned 48 hours later to complete the second session. Occasionally, participants could not return exactly 48 hours later.

When participants returned for Session 2 they completed the recall task again, followed by retention trials. During retention trials the keypress sequence was presented without the goal time and no feedback was given. The keypress sequence was used as the stimuli so that the retention task focused performance on the critical task characteristic: producing the overall movement according to the goal time. Participants were instructed that their goal was still to complete the movement according to the goal times specified during Session 1 training. Participants completed two retention blocks (12 trials per block).

One of the retention blocks was ordered using a blocked schedule and the other retention block was ordered using a random schedule. The order of the blocked and random retention blocks were counterbalanced over participants. The blocked block presented three consecutive trials of each movement pattern; the order of the movement patterns matched the order of patterns in that participant's acquisition schedule version (e.g., Blocked–1 or Blocked–2). The random block used the first random block from each participant's respective acquisition schedule version (e.g., Random–1, Block 1 or Random–2, Block 1). Because acquisition was manipulated between subjects, the random or blocked retention block was selected by matching participants' version variable (e.g., the Random–1 acquisition schedule condition was assigned to a random retention block that matched the first acquisition block from Random–1 schedule version and the blocked retention block used the order of movement patterns from the Block–1 schedule version).

Participants next completed two blocks of a *double pattern* transfer task. In the double pattern task participants had to consecutively perform two of the movement patterns that they had previously learned, without pausing during the transition from Pattern A to Pattern B. The double pattern transfer task has two movement patterns, either the *same* patterns or *different* patterns. An example of a same pattern is *1500–1500*, where participants have to perform the 1500 ms pattern twice. An example of a different sequence is *1500–900*, where participants have to perform the 1500 ms pattern and then the 900 ms pattern. Across the two blocks there were 8 same sequence trials and 16 different sequence trials.

During double pattern task trials participants saw only the two goal movement times, no keypress sequence diagrams were shown. As a result participants had to recall the keypress sequences based on the movement times and perform each according to its associated goal time. After each trial participants received limited feedback. If their keypress sequence was incorrect then the correct sequence (or sequences) was displayed. If both keypress sequences were correct, then the movement time for each pattern and the total movement time was displayed; additionally the differences between the three goal times (Pattern A, Pattern B, both patterns) and the participant's actual movement times were displayed. If either of the keypress sequences was incorrect the trial repeated immediately.

I hypothesized that if a random acquisition schedule improves fluency with loading and preparing action plans (e.g., Lee & Magill, 1983) and sequencing responses in working memory (e.g., Carlson & Shin, 1996), then participants in the random schedule condition should be more facile at retrieving and preparing the motor programs needed to create the novel (double) action plan. This could result in faster times to prepare the novel action plan and fewer errors in executing the plan.

Lastly, participants completed two blocks of a *mirror pattern* transfer task. In the mirror pattern task participants had to perform the previously learned movement patterns by mirror-reversing them. Participants were shown a goal time and they had to produce a keypress sequence by recalling the original movement pattern and then mirroring it along its vertical axis. Table 7 lists the keypress sequences for these mirrored plays. During mirror pattern transfer trials participants saw only the goal movement time, they did not see the goal keypress sequence. If participants performed an incorrect keypress sequence then their feedback included the original

Table 7: The four movement patterns (keypress sequence and goal movement time) used during the acquisition and mirror transfer tasks in Experiments 1 and 2.

Phase	Goal time	Keypress sequence
Acquisition	900 ms	9–5–1–2–3
Acquisition	1050 ms	2–1–6–4–8
Acquisition	1200 ms	3–6–5–8–4
Acquisition	1500 ms	4–2–5–8–9
Transfer	900 ms (mirror)	7–5–3–2–1
Transfer	1050 ms (mirror)	2–3–4–6–8
Transfer	1200 ms (mirror)	1–4–5–8–6
Transfer	1500 ms (mirror)	6–2–5–8–7

(not mirrored) keypress sequence. Incorrect trials were repeated immediately. If participants performed the keypress sequence correctly then they were given the same feedback as during acquisition: (1) accuracy of the keypress sequence (correct or incorrect), (2) duration of the keypress sequence, and (3) difference between their movement time and the goal time. All blocks were ordered using a random schedule.

I hypothesized that the mirror transfer task would impose a working memory load because participants have to use the goal time to retrieve the keypress sequence, mentally rotate it, and then create a novel action plan (or modify an existing action plan). If the random schedule condition does improve the strength of the representation of the movement pattern, then participants should be better able to retrieve, modify, and program the action plan. This might result in less planning that must be done during response execution, thereby reducing timing errors (Klapp & Wyatt, 1976).

2.3 Results

A core analysis framework was used for Experiments 1–4. All analysis were interpreted with $\alpha = .05$. When repeated-measures analyses were conducted, they were conducted using a multivariate approach (i.e., a MANOVA framework; O'Brien & Kaiser, 1985). In MANOVA and ANOVA analyses, "Type II" Sums of Squares was used, which conforms to the principle of marginality (Nelder, 1977) and is appropriate for unbalanced designs (Fox & Weisberg, 2011). When not planned a priori, follow-up and pairwise comparisons used the Tukey HSD procedure[2]. Log transformations used the natural log.

2.3.1 Structuring acquisition blocks

In order to compare performance on equivalent trials, data in the blocked and blocked-repeated schedules were restructured following established procedures in the practice schedule literature (see Lee & Magill, 1983; J. B. Shea & Morgan, 1979). For data analysis, I created new acquisition blocks that included three trials for each of the four movement patterns (12 trials total). Thus, the first acquisition block contained the first, second, and third trials that a participant correctly performed for each movement pattern, regardless of where those trials occurred during training. The second acquisition block contained the fourth, fifth, and sixth trials that a participant correctly performed for each movement pattern. I repeated this procedure for the remaining 34 blocks. Creating these blocks required restructuring the data for the blocked and blocked-repeated groups; no restructuring was needed for the random groups because their blocks already had three trials for each of the four movement patterns.

2.3.2 Dependent Measure

2.3.2.1 Total error

Given that the goal of the multisegment movement task is to produce the movement sequence according to a goal time, any variability around this goal time is considered error. *Root mean square error (RMSE)*, sometimes referred to as *total error*, indicates both the bias and variability of the timing of responses for each participant (Schmidt

[2]Tukey's HSD controls α family-wise while retaining power and is appropriate for balanced data with equal error variance. The TukeyHSD function in the R `stats` package was used to implement Tukey's HSD method, and includes an adjustment for unbalanced designs.

& Lee, 2005). When multiple trials exist for a given movement pattern, RMSE indicates timing errors as it measures the variability around the goal movement time for the set of trials

$$\text{RMSE} = \sqrt{\frac{\Sigma(x_i - T)^2}{n}} \tag{1}$$

where x_i is movement time on trial i, T is the goal time, and n is the number of trials the participant performed (Schmidt & Lee, 2005). Two of the three amount of practice studies in the CI domain have used RMSE as the primary dependent measure (see Table 1). A further discussion of RMSE as a dependent measure is included in Appendix B.

Total error was computed for correct trials. Trials were considered correct if all the keys were pressed in the correct order.

2.3.2.2 Response Time

One of the transfer tasks also uses *response time (RT)*, sometimes referred to as *response latency*, as a dependent measure. Response time is the time between the presentation of a stimulus and the execution of a response. Response time is hypothesized to reflect mental processing.

2.3.3 Removing Outlier Trials

A trimming procedure was used to address extreme outlier trials in Experiments 1 and 2. Prior to analysis, movement times (and response times, when used) for correct trials were analyzed to identify and remove outliers. These outliers were almost always in the upper tail of the distribution and were assumed to arise from an external distraction during the trial. Movement times were trimmed if they were more than six standard deviations from the mean (in Session 1) or four standard deviations from the mean (in Session 2). The higher standard deviation cutoff in the acquisition session was used to be overly conservative in removing data, as early acquisition likely has more variable times due to participants' inexperience with the task. Trimming was always done separately for each epoch. Additionally, trimming was performed separately for each movement pattern because movement patterns have different mean movement times.

2.3.4 Session 1

In this multisegment movement task, the random acquisition schedule is hypothesized to increase the difficulty of the task during acquisition. On the majority of trials a participant has to prepare to execute a different movement pattern than on the previous trial. This additional preparation is hypothesized to result in more difficulty planning, and thus higher RMSE rates. However, this practice preparing movement patterns is also hypothesized to result in more durable storage and efficient retrieval and therefore lower RMSE during retention trials and transfer trials.

2.3.4.1 RMSE During Acquisition

After restructuring the acquisition data RMSE was calculated for each block. Figure 11 shows RMSE during acquisition for the three practice schedules. In general, the data demonstrate the law of practice: performance improves (i.e., RMSE decreases) rapidly early in training and less rapidly later in training. Additionally, the data suggest the practice schedule manipulation is affecting RMSE. On the initial block of practice there is a separation between the random acquisition schedule and the blocked and blocked-repeated acquisition schedules. In the final block of practice there is a separation between the random and blocked schedules; the blocked-repeated schedule is between the other two schedules. In the bulk of the middle blocks the data is more variable; at some blocks the schedules do not seem to differ whereas in other blocks the schedules do seem to differ. In order to more specifically examine these trends, I conducted a two-step analysis. First, I tested the effect of practice schedule on RMSE during the initial acquisition block (Block 1) and the final acquisition block (Block 36). Second, to characterize the effect of practice schedule on RMSE during

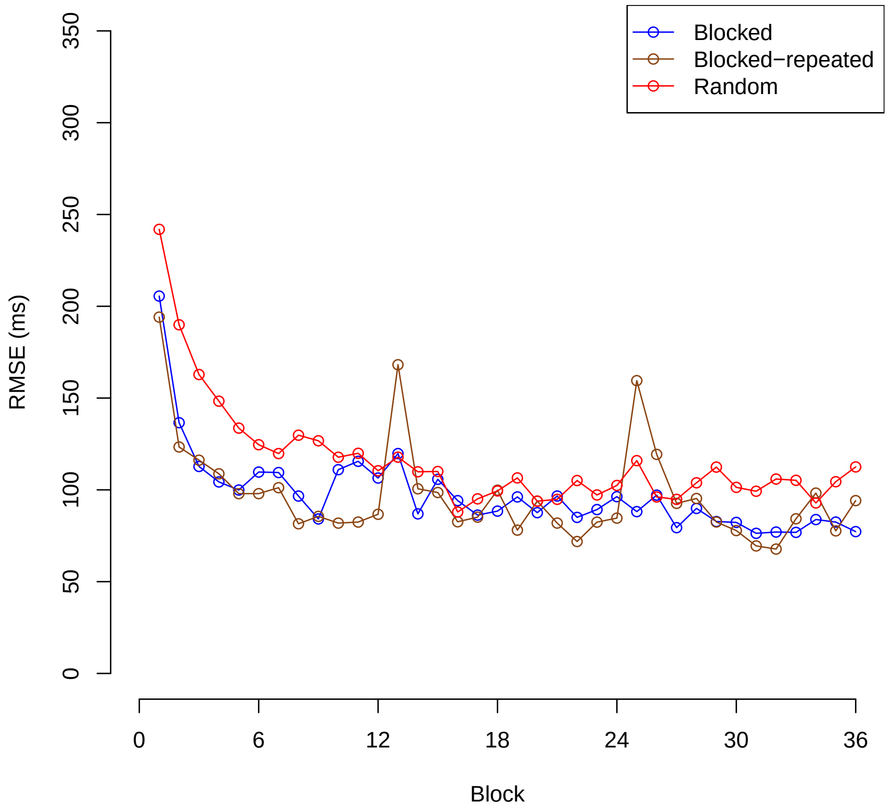

Figure 11: RMSE during acquisition blocks, as a function of practice schedule.

the entirety of acquisition, I fit the aggregate data from the three practice schedules to curves using a simple power law.

Performance on the initial and final blocks. RMSE on the initial and final blocks was analyzed using two univariate ANOVAs with practice schedule as a between subject factor. Practice schedule did not have a significant effect on RMSE during the initial block, $F_{(2,34)} = 1.34$, $p = .28$, $RMSE = 5626.00$. Additionally, practice schedule did not have a significant effect on RMSE during the final block, $F_{(2,34)} = 2.93$, $p = .07$, $MSE = 1324.77$. The absence of statistically significant differences between practice schedules during the initial and final blocks is counter to the visual trends displayed in Figure 11. This suggests a more appropriate analysis is

Table 8: Experiment 1. Model fit and parameters derived from fitting each practice schedule to a simple power law.

		Parameters	
Acquisition schedule	R^2	β	α
Blocked	.72	254.22	.19
Blocked-repeated	.29	199.70	.15
Random	.80	341.79	.21

to model RMSE over the duration of acquisition training, for each practice schedule. To do this I fit the empirical data to learning curves derived from a mathematical model of a simple power law.

Fitting Learning Curves RMSE for each practice schedule was fit to a simple power curve

$$Y = B + N^{-\alpha} \tag{2}$$

where Y is performance (RMSE), N is trial number, B is initial performance, and α is the learning rate parameter (Newell & Rosenbloom, 1981).

Model fit and learning curve parameters are displayed in Table 8. Figure 12 displays these learning curves along with the empirical data to which they are fit. The learning rate parameter (α) and the initial performance parameter (β) is highest for the random schedule, and lowest for the blocked-repeated schedule. As can be seen in Figure 12, the initial level of RMSE for the random schedule is greater than the other schedules (more than 100 ms greater than both the blocked and blocked-repeated group). Although the random schedule has a larger learning rate parameter, it is not large enough for RMSE to decrease enough to the difference in initial starting points between the blocked and random schedules. These data suggest that the random schedule creates a context in which initial performance is worse than performance under a blocked schedule and, at the end of training, performance in the random schedule remains worse than the blocked schedule.

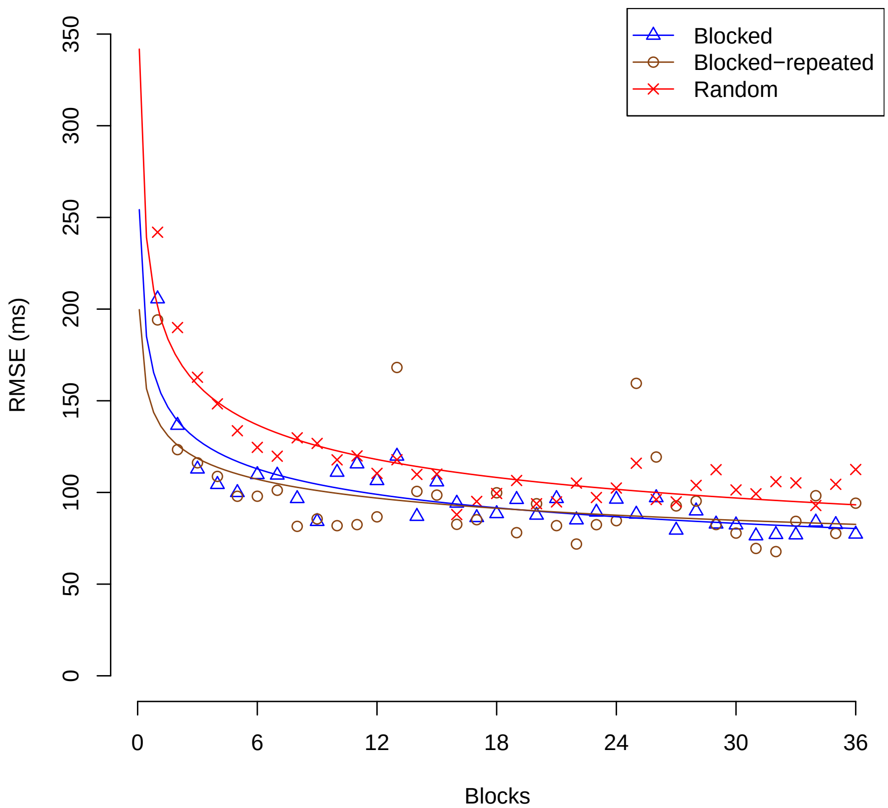

Figure 12: RMSE during acquisition, with curves based on fitting data to a simple power law, as a function of practice schedule.

Performance in the blocked-repeated schedule closely matches performance in the blocked schedule, especially by the end of Epoch 1 (Block 12). Of the three practice schedules, the blocked-repeated schedule has the worst fit. One cause of the ill fit is the spike in RMSE at Block 13 and Block 25 (and, to a lesser extent, Block 14 and Block 26). This spike may be thought of as a *task switch cost* and appears in these blocks because of the distribution of tasks (movement pattern trials) in the blocked-repeated schedule. In the random schedule, tasks switched at least once every three trials (i.e., several times within each block). In the blocked schedule, tasks switched only four times during all of acquisition: once for each movement pattern. In the blocked-repeated schedule, tasks switched 12 times during acquisition, three times for each movement pattern (because the blocked sequence of four movement patterns is repeated three times).

Because of how the acquisition blocks are restructured for data analysis, trials with a task switch appear in different blocks for each practice schedule. In the random schedule trials with a task switch occur in *each block*. In the blocked schedule, trials with a task switch occur in *Block 1 only*. In the blocked-repeated schedule, trials with a task switch occur in *Blocks 1, 13, and 25*.

Therefore, if one assumes task switches have a cost (e.g., higher RMSE), those costs would be distributed throughout all of the acquisition blocks in the random schedule, but concentrated in Block 1 for the blocked schedule and Blocks 1, 13, and 25 for the blocked-repeated schedule.

These data suggest that the timing of task switches in practice schedules is one example of the contextual variety (Battig, 1972) that might create additional processing demands on performance. According to the new theory of disuse (Bjork & Bjork, 1992), these numerous, repeated retrievals in the random schedule might increase retrieval strength. Moreover, the less frequent retrievals in the blocked schedule might increase storage strength. This processing account is consistent with the retention

Table 9: Experiment 1. Number of participants (during Session 2, retention) in each acquisition schedule and retention order cell.

	Retention order		
Acquisition schedule	Blocked first	Random first	Total
Blocked	8	5	13
Blocked-repeated	7	5	12
Random	8	4	12
Total	23	14	37

effects found in Session 2.

2.3.5 Session 2

2.3.5.1 Skill Retention

Data Analysis Procedure The data analysis procedure followed the acquisition procedure. Movement times that were more than three standard deviations away from the mean (for each task) were trimmed. Data from 12 trials were trimmed (12 of 888 total trials were trimmed, 1.35% of the retention data).

Participants completed two retention blocks: a blocked retention block and a random retention block. The order of the retention blocks was counterbalanced across participants. This counterbalance variable, *retention order* was included in the statistical models, in order to account for any differences that might arise from doing the blocked retention block first or second.

Due to removing some participants, the counterbalance was not completely balanced. Cell sizes for each condition are shown in Table 9.

RMSE RMSE was calculated separately for the blocked and random retention blocks. RMSE descriptive statistics are shown in Table **??**. RMSE was analyzed separately for each retention block, using two-way ANOVAs with acquisition schedule and retention order as between subject factors.

In the blocked retention block, there was no significant difference among acquisition schedules, $F_{(2,31)} = 1.34, p = .28, MSE = 7330.77$. There was no significant effect of retention order, $F_{(1,31)} = 1.34, p = .26$, nor was there a significant interaction between acquisition schedule and retention order, $F_{(2,31)} = 1.14, p = .33$.

In the random retention block, there was a significant difference among acquisition schedules, $F_{(2,31)} = 3.71, p < .05, MSE = 8022.26$. There was no significant effect of retention order, $F_{(1,31)} = 0.49, p = .49$, nor was there a significant interaction between acquisition schedule and retention order, $F_{(2,31)} = 0.44, p = .65$. The significant schedule main effect was followed by pairwise comparisons using Tukey's HSD procedure. The blocked schedule had significantly higher RMSE than the random group (mean difference = 92.21 ms, $p < .05$) but did not have a significantly different RMSE than the blocked-repeated group (mean difference = 11.86 ms, $p = .94$). The random schedule had lower RMSE than the blocked-repeated group, but this difference was not significant (mean difference = 80.35 ms, $p = .08$).

2.3.5.2 Pattern Recall

Participants completed the paper and pencil recall measure twice: at the end of Session 1 and the beginning of Session 2. A repeated measures analysis was conducted using acquisition schedule as a between subject factor and session as a repeated measure. Acquisition schedule had a significant effect on pattern recall (see Table 10), $V = 0.28$, $F_{(2,33)} = 6.50$, $p < .005$. There was no significant session effect, $V = 0.04$, $F_{(1,33)} = 1.36$, $p = .25$, nor was there a significant schedule X session interaction, $V = 0.03$, $F_{(2,33)} = 0.43$, $p < .65$.

The significant schedule main effect was followed by pairwise comparisons. The blocked schedule had significantly lower recall than the blocked-repeated group (mean difference = 0.31, $p < .05$), and the random group (mean difference = 0.33, $p < .01$). There was no significant difference in recall between the blocked-repeated group and

Table 10: Experiment 1. Mean proportional recall accuracy (and standard deviation) as a function of acquisition schedule and session.

	Session		
Acquisition schedule	1	2	Average
Blocked	0.38 (0.24)	0.38 (0.28)	0.38 $(0.19)_a$
Blocked-repeated	0.75 (0.28)	0.65 (0.29)	0.70 $(0.26)_b$
Random	0.77 (0.36)	0.68 (0.30)	0.72 $(0.32)_b$

Note. One participant in the random schedule condition did not complete the retention measure in Session 2. This participant was excluded from the Session 2 and Average statistics.
Means in each column that share a subscript do not differ significantly ($p < .05$).

the random group (mean difference = 0.02, $p = .98$).

2.3.6 Double Pattern Transfer Task

The double pattern transfer task was analyzed for differences among acquisition practice schedules. One additional participant (blocked-repeated condition) was dropped from the double pattern transfer task because of a software error during the task. Group sizes were unequal: blocked schedule condition = 13, blocked-repeated schedule condition = 11, random schedule condition = 12.

Two dependent variables were analyzed: errors and response time. Errors were counted by summing errors on the first pattern (Pattern A) and errors on the second pattern (Pattern B). For each trial, a participant could have between zero and two errors. Errors were hypothesized to indicate an inability to recall the movement pattern, difficulty in executing the pattern, or difficulty in creating a novel motor action plan from the two previously learned movement patterns.

Response time equals the time elapsed between displaying the stimulus (the goal movement time of the two patterns) and the first keypress. Response time is hypothesized to reflect planning time, the time in which participants process the stimuli and plan the upcoming movement, such as retrieving, modifying, and parameterizing an action plan (or plans).

RMSE was not analyzed because the design of the task precludes meaningful interpretation of RMSE. The double task was designed to encourage participants to plan the double movement prior to beginning a movement (i.e., not to execute the first keypress sequence, pause, plan the second keypress sequence, then execute the sequence). To encourage this pre-planning, the software continues counting elapsed time during the interval between pressing the 5th key and the 6th key. Thus, if participants are executing both patterns at precisely their goal time, they still need to try to minimize the time between pressing the 5th key and the 6th key. Only by minimizing this time can one get close to executing the double pattern according to the *overall goal movement time*. The feedback given to participants illustrates the timing of each pattern and total time, signalling to participants that any pause between patterns contributes to overall time. Nevertheless, participants have to move between the 5th and 6th key which necessarily increases their movement time above the goal time. Participants can adopt multiple strategies to try to optimize different movement times (i.e., for Pattern A, for Pattern B, or for the entire pattern: A + B), but there is a tradeoff such that it is impossible to simultaneously optimize all three. Therefore, there is a built in positive bias in any measure of RMSE.

2.3.6.1 Errors

Errors were analyzed using a repeated measures analysis with acquisition schedule as a between-subject factor and block as a repeated measure. Block refers to the first or second block of trials on the double pattern transfer task. Mean errors (and standard deviations) for each acquisition schedule condition and transfer block are displayed in Table 11. Block had a significant effect on errors (see Table 11), $V = 0.37$, $F_{(1,33)} = 6.6$, $p < .05$, with errors decreasing from Block 1 to Block 2. Acquisition schedule did not have a significant effect on errors, $V = 0.05, F_{(2,33)} = 0.9, p = .41$, nor

Table 11: Experiment 1. Mean errors (and standard deviation) on the double pattern task as a function of acquisition schedule and block.

	Block		
Acquisition Schedule	1	2	Average
Blocked	8.00 (6.68)	2.54 (3.02)	5.27 (4.73)
Blocked-repeated	6.55 (7.37)	2.45 (2.42)	4.50 (4.28)
Random	4.08 (3.94)	2.17 (2.08)	3.13 (2.24)

was there a significant interaction between acquisition schedule and transfer block, $V = 0.11$, $F_{(2,33)} = 2.0$, $p = .15$.

2.3.6.2 Response Time

Response times for correct trials were calculated for each of the double pattern transfer task blocks. Median response times for each block were used to reduce the effect of outlier trials. Mean median response times (and standard deviations) for each acquisition schedule condition and transfer block are displayed in Table 12.

Prior to testing the effects of acquisition schedule and block, median response times were log transformed to normalize their distributions. A repeated measures analysis was conducted with acquisition schedule as a between subject factor and block as a repeated measure. Block had a significant effect on response times (see Table 12), $V = 0.37$, $F_{(1,33)} = 19.61$, $p < .001$, with response times decreasing from Block 1 to Block 2. Acquisition schedule did not have a significant effect on response time, $V = 0.05$, $F_{(2,33)} = 0.95$, $p = .40$, nor was there a significant interaction between acquisition schedule and transfer block, $V = 0.08$, $F_{(2,33)} = 1.45$, $p = .25$.

2.3.7 Mirror Transfer Task

The mirror transfer task was analyzed for differences between acquisition schedules. Group sizes were unequal, blocked condition = 13, blocked-repeated condition = 12, random condition = 12.

Table 12: Experiment 1. Mean median response time (and standard deviation) on the double pattern task as a function of acquisition schedule and block.

	Block		
Acquisition Schedule	1	2	Average
Blocked	4640.08 (1802.41)	4160.12 (1474.05)	4400.10 (1478.83)
Blocked-repeated	4033.91 (2651.99)	3705.09 (2387.33)	3869.50 (2473.64)
Random	4700.83 (1744.39)	3398.75 (1175.51)	4049.79 (1402.24)

Table 13: Experiment 1. Mean RMSE (and standard deviation) on the mirror transfer task, as a function of acquisition schedule and block.

	Block		
Acquisition Schedule	1	2	Average
Blocked	240.49 (82.28)	241.68 (108.29)	241.09 (91.86)
Blocked-repeated	221.38 (68.25)	242.20 (71.37)	231.79 (62.90)
Random	173.70 (84.58)	161.86 (99.04)	167.78 (86.61)

Data analysis followed the procedure for acquisition and retention data. Movement times that were more than three standard deviations away from the mean (for each task) were trimmed. Data from 12 trials were trimmed (12 of 888 total trials were trimmed, 1.35% of the mirror transfer data).

Participants completed two mirror transfer task blocks. RMSE was calculated for each block. RMSE was analyzed using a repeated measures analysis with acquisition schedule as a between subject factor and block as a repeated measure. RMSE descriptive statistics are shown in Table 13.

As with retention, the blocked condition has a higher RMSE than the random condition, but the acquisition schedule effect was not significant, $V = 0.15$, $F_{(2,34)} = 2.91$, $p = .07$. There was no significant effect of block, $V < 0.01$, $F_{(1,34)} = 0.11$, $p = .74$, nor was there a significant acquisition schedule X block interaction, $V = 0.05$, $F_{(2,34)} = 0.90$, $p = .42$.

2.4 Discussion

As expected, over the course of acquisition training, the blocked schedule resulted in lower RMSE than the random schedule. Fitting the learning curves showed large differences in the initial rate of performance between the blocked (and blocked-repeated) and the random schedule condition. This performance difference between schedules persisted even at the end of acquisition training.

Unexpectedly, the blocked-repeated schedule showed a different performance profile than the blocked schedule. In particular, the blocked-repeated schedule showed spikes in RMSE during blocks immediately following a change to a new movement pattern.

Acquisition practice schedule affected both measures of retention. Pattern recall was lower in the blocked schedule condition than in either the blocked or random schedule conditions. Skill retention showed the same pattern of practice schedule differences, but only in the retention block where trials were ordered using a random schedule. In this random retention block the blocked acquisition schedule results in higher RMSE than either the blocked-repeated or random schedule conditions. For the blocked and blocked-repeated schedule this random retention block represents a test of retention plus transfer of practice schedule.

The transfer measures showed no significant effects of practice schedule, although there was a trend for lower RMSE in the random schedule on the transfer mirror task.

2.5 Implications for Experiment 2

These results suggest it would be inappropriate to use a blocked-repeated schedule as a substitute for a blocked schedule. A blocked-repeated schedule should be investigated more extensively, as it seems to provide some of the benefits of a blocked schedule during acquisition and some of the benefits of the random schedule during retention. However, the results show that a blocked-repeated schedule is not simply a variation of a blocked schedule that can be implemented when training with high amounts of practice. Instead, the blocked-repeated schedule is a distinct type of practice schedule.

The results also inform the selection of the number of trials to use to operationalize the different amounts of practice in Experiment 2. The learning curves provided a visual guide for identifying where performance is still rapidly improving (low amount), where performance is beginning to level off (medium amount), and where performance improvement is slowing even further (high amount).

CHAPTER III

EXPERIMENT 2

Experiment 2 was conducted to evaluate the effects of practice schedule and amount of practice on acquisition, retention, and transfer of the multisegment movement task.

3.1 Method

3.1.1 Participants

One-hundred and twenty-five participants over the age of 17 were recruited from the Georgia Tech participant pool. Nineteen participants were excluded prior to data analysis for two reasons: ignored instructions on how to perform the task (N = 17)[1] or continuously talked and acted disruptive during the session (N = 2).

Excluding these participants results in 106 participants; cell sizes are displayed in Table 14. Participants' mean age was 19.4 years ($SD = 1.5$). Of the 106 participants, 57 were male, 47 were female, and 2 did not answer. Participants self-reported their dominant hand: 90 were right-handed, 15 were left-handed, and 1 was ambidextrous.

3.1.2 Design

3.1.2.1 Practice schedule

Two practice schedules were used: blocked, and random. Experiment 2 used the same blocked and random schedules (and schedule versions) that were used in Experiment 1.

[1]Participants were considered to have ignored instructions if they were performing the movement patterns as fast as possible as indicated by either (a) a note from the experimenter in the subject log or (b) their acquisition data. Participants' data was dropped if their data met the exclusion criteria that constant error (CE) was less than -200 ms on $\frac{1}{3}$ or more of their acquisition blocks. A consistent CE value in this range or lower indicates that participants were continuously performing the movement patterns as fast as possible, rather than according to the goal movement time.

Table 14: Experiment 2. Number of participants (during Session 1, acquisition) in each acquisition schedule, acquisition schedule version, and amount of practice cell.

	Version		
Acquisition Schedule	1	2	Total
Low Amount			
Blocked	8	9	17
Random	9	11	20
Total	17	20	37
Medium Amount			
Blocked	8	9	17
Random	9	8	17
Total	17	17	34
High Amount			
Blocked	8	10	18
Random	10	7	17
Total	18	17	35
Total			
Blocked	24	28	52
Random	28	26	54
Total	52	54	106

3.1.2.2 Amount of practice

Participants were assigned to one of three amount of practice conditions: low, medium, or high. The low amount had 3 blocks (36 trials); the medium amount had 12 blocks (144 trials); the high amount had 24 blocks (288 trials). These amounts were selected based on the learning curves fit from the Experiment 1 data, comparisons to existing research using the practice schedule manipulations, and the practical constraints of creating blocks of trials that could be equally sequenced with four task categories.

Specifically, the low amount was chosen to correspond to a location where there was still rapid improvement in learning and a large separation between performance in the blocked and random schedule conditions. The medium condition was placed where performance improvement began leveling off. Additionally, blocks were chosen such that positive and negative examples could be distributed consistently across blocks in all low, medium, and high amount conditions. Finally, an effort was made to match relative differences between amounts of practice to existing research. The low amount of practice used 25% of the trials in the medium amount of practice condition and the high amount of practice used 200% of the trials in the medium amount of practice condition; these percentages match those used by C. H. Shea et al. (1990), which showed an interaction between practice schedule and amount of practice.

3.1.2.3 Retention interval

As in Experiment 1, participants completed two sessions: Session 1 (acquisition), and 48 hours later, Session 2 (retention and transfer).

3.1.3 Materials

The same multisegment movement tasks and stimuli from Experiment 1 were used.

3.1.4 Procedure

The same procedure as Experiment 1 was used. The only change to the procedure resulted from the amount of practice manipulation. In Experiment 1 participants completed paper and pencil tasks during the breaks between epochs, in order to minimize fatigue effects. In Experiment 1 an epoch was defined as 12 blocks. Experiment 2 epochs also had 12 blocks, which means that only the high amount condition had more than one epoch. Regardless of condition, however, all participants completed the same paper and pencil tasks. In the low amount and medium amount conditions participants completed these tasks after their final acquisition block (Blocks 3 and 12, respectively). In the high amount condition participants completed the first set of tasks after Epoch 1 (Block 12) and the second set after Epoch 2 (Block 24).

3.2 Results

3.2.1 Session 1

As in Experiment 1, movement times that were more than six standard deviations from the mean were trimmed. Trimming was done separately for each amount of practice condition. In the high amount of practice condition, trimming was done separately for each epoch. In total, 27 values were trimmed (low: 2 of 1332, 0.15%; medium: 8 of 4896, 0.16%; high: 17 of 20160, 0.08%).

RMSE was calculated for each block, following the procedure for restructuring blocks described in Experiment 1. RMSE during acquisition data is graphed for the low, medium, and high amount conditions in Figures 13, 14, and 15, respectively. Figure 16 shows RMSE during acquisition for all practice schedule and amount of practice conditions.

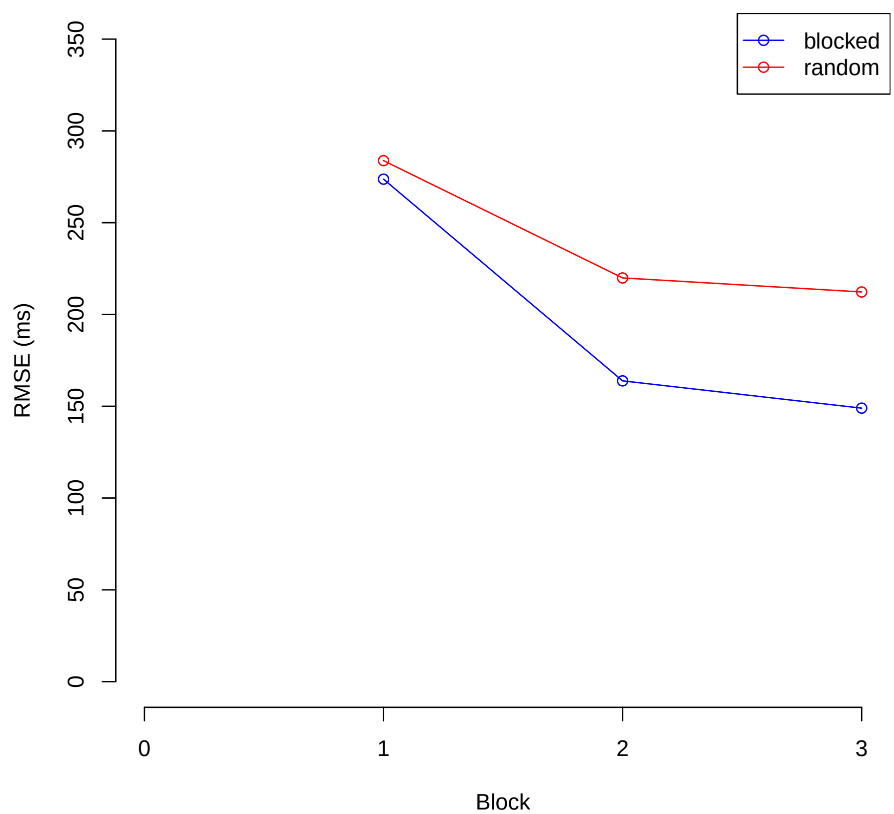

Figure 13: RMSE during acquisition blocks for the low amount of practice condition, as a function of practice schedule.

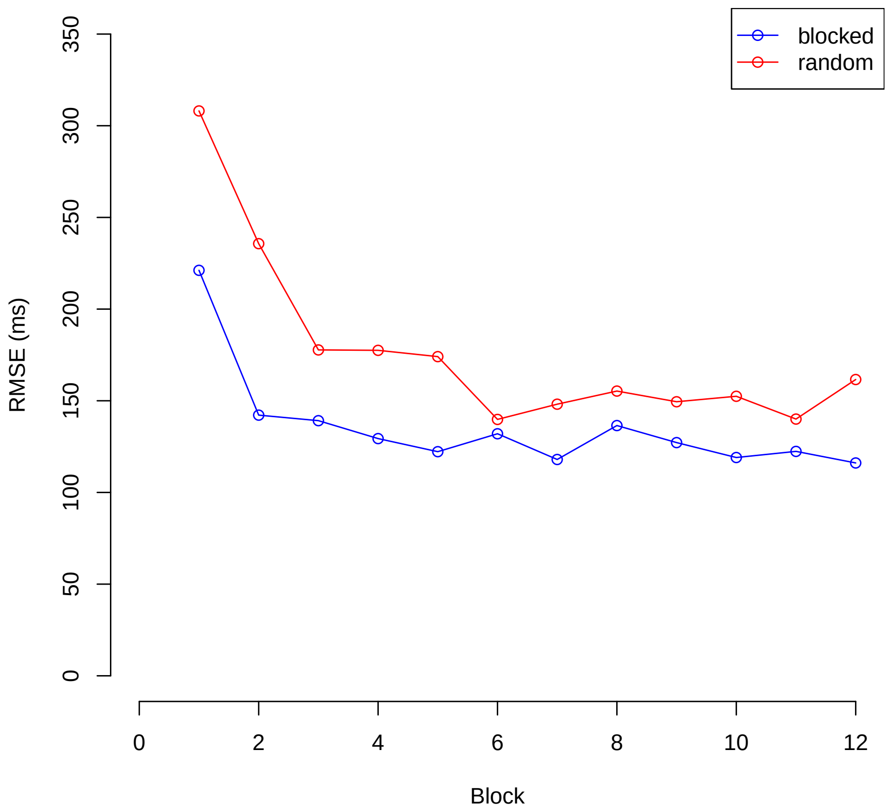

Figure 14: RMSE during acquisition blocks for the medium amount of practice condition, as a function of practice schedule.

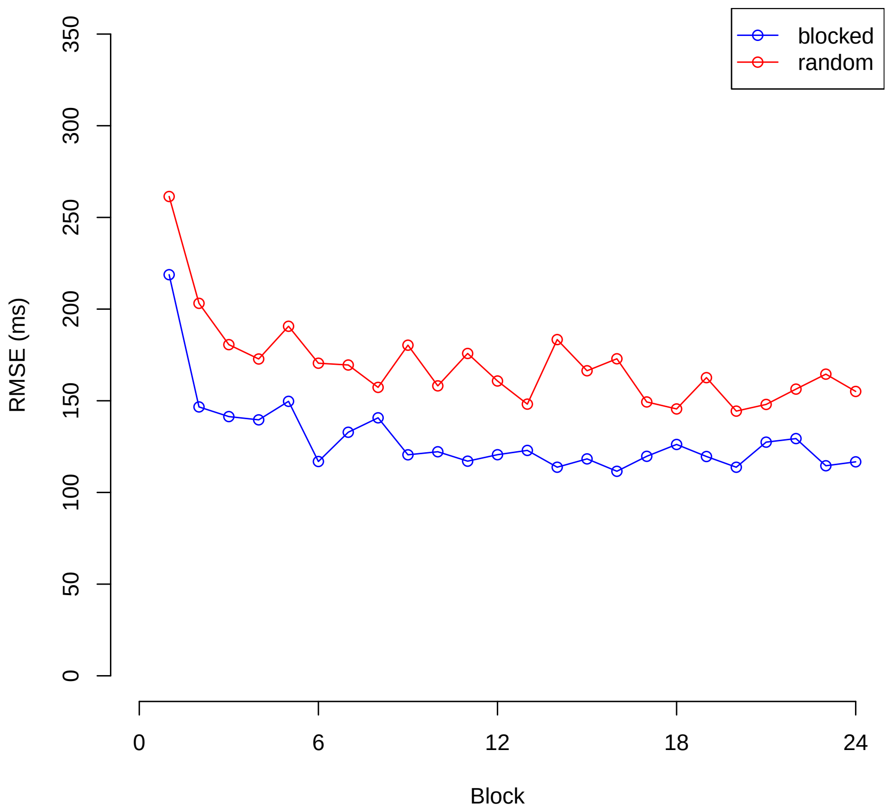

Figure 15: RMSE during acquisition blocks for the high amount of practice condition, as a function of practice schedule.

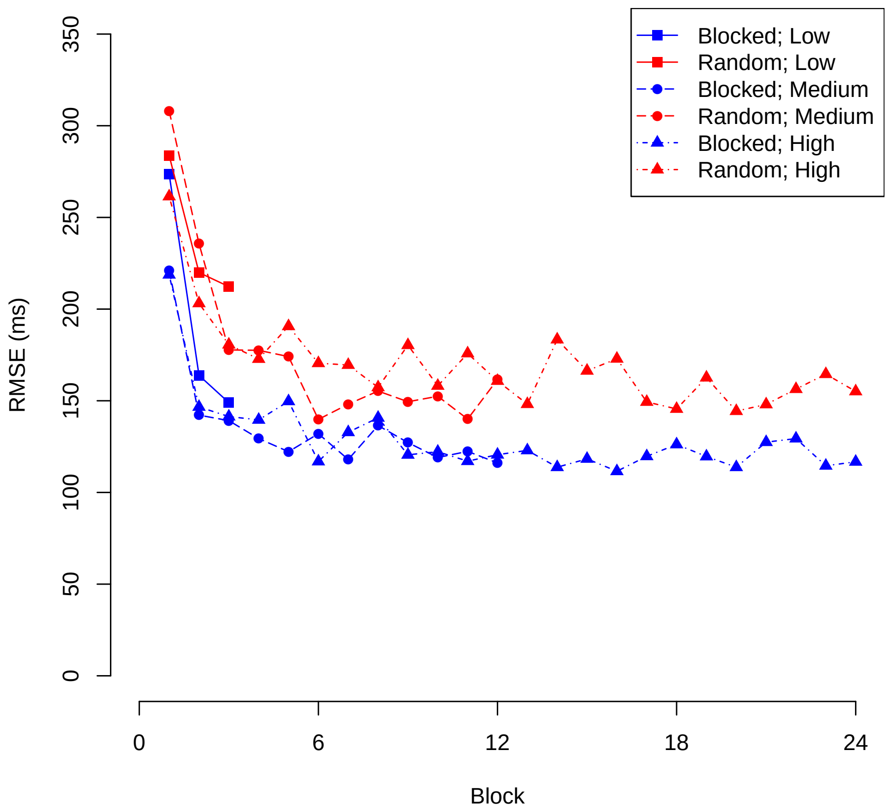

Figure 16: RMSE during acquisition blocks, as a function of practice schedule and amount of practice.

3.2.1.1 Performance on the Final Block

To assess the effect that acquisition schedule and amount of practice have at the end of training, I examined the performance during the final acquisition block. Means and standard deviations for each acquisition schedule by amount of practice condition are displayed in Table 15. The distribution of RMSE values in this final acquisition block had a positive skew, resulting in a departure from normality. Prior to statistical analysis, RMSE was log-transformed, which produced a normal distribution of RMSE values. Performance on the final acquisition block (log RMSE) was modeled using a full factorial model with acquisition schedule and amount of practice.

There was a significant effect of acquisition schedule, $F_{(1,100)} = 13.52$, $p < .001$, $MSE = 0.22$ with the blocked schedule resulting in lower RMSE than the random schedule (see Table 15). Additionally there was a significant effect of amount of practice, $F_{(2,100)} = 5.19$, $p < .01$. There was no significant higher order interaction, $F_{(2,100)} = 0.04$, $p = .96$.

The significant amount of practice effect was followed with pairwise comparisons using the Tukey HSD procedure. The medium amount of practice condition had significantly lower RMSE than the low amount of practice condition (mean difference of log RMSE = 0.29, $p < .05$). Likewise, the high amount of practice condition had significantly lower RMSE the low amount of practice condition (mean difference of log RMSE = 0.35, $p < .01$). There was no significant difference in RMSE between the medium amount and high amount of practice conditions (mean difference of log RMSE = -0.07, $p = 0.84$). These results indicate that increasing practice from the low to medium amount decreased RMSE, but continued increase of practice from medium to high amount did not significantly decrease RMSE. This does not necessarily mean that RMSE reached asymptote by the end of the medium amount of practice. Instead, RMSE might still be decreasing, but the increased practice from 12 blocks (medium amount) to 24 blocks (high amount) was not substantial enough to cause a significant decrease in RMSE.

3.2.2 Session 2

Nine participants did not return for Session 2 and were excluded from further analysis. Cell means for the retention and recall measures are displayed in see Table 16.

3.2.2.1 Skill Retention

Data Analysis Procedure The data analysis procedure followed the acquisition procedure. Movement times that were more than four standard deviations away from the mean were trimmed. Data from four trials were trimmed (4 of 2328 total trials were trimmed, 0.17% of the retention data).

RMSE RMSE was calculated separately for the blocked and random retention blocks. RMSE descriptive statistics for the blocked retention block and the random retention block are shown in Tables 17 and 18, respectively.

RMSE was analyzed using an ANOVA with acquisition schedule, amount of practice, and retention order as factors. As in Experiment 1, retention blocks was counterbalanced across participants and this counterbalance variable (retention order) was included in the statistical models. The models tested the main effects of each factor and the two-way interactions between acquisition schedule and amount of practice, and between acquisition schedule and retention order.

In the blocked retention block, there was a significant difference between acquisition schedule, $F_{(1,89)} = 5.47, p < .05, MSE = 7534.06$; on average the random schedule condition had 40 ms lower RMSE than the blocked schedule condition (see Table 17). There was also a significant difference between amount of practice, $F_{(2,89)} = 3.10, p < .05$. There was no significant interaction between acquisition schedule and amount of practice, $F_{(2,89)} = 0.82, p = .44$, nor was there a significant interaction between acquisition schedule and retention order $F_{(1,89)} = 0.03, p = .87$.

Table 15: Experiment 2. Mean RMSE (and standard deviation) during the final acquisition block, as a function of acquisition schedule and amount of practice.

	Amount			
Acquisition schedule	Low	Medium	High	Average
Blocked	148.99 (66.45)	116.09 (52.21)	116.71 (76.30)	127.06 $(66.48)_a$
Random	212.27 (87.32)	161.59 (70.98)	155.10 (80.62)	178.32 $(83.15)_b$
Average	183.19 $(83.72)_a$	138.84 $(65.56)_b$	135.35 $(79.68)_b$	153.17 (79.37)

Note. In the Average row and the Average column, means that share a subscript do not differ significantly ($p < .05$).

Table 16: Experiment 2. Number of participants (during Session 2, retention) in each acquisition schedule, amount of practice, and retention order cell.

	Retention order		
Acquisition schedule	Blocked first	Random first	Total
	Low Amount		
Blocked	7	8	15
Random	9	8	17
Total	16	16	32
	Medium Amount		
Blocked	9	7	16
Random	8	7	15
Total	17	14	31
	High Amount		
Blocked	9	9	18
Random	8	8	16
Total	17	17	34
	Total		
Blocked	25	24	49
Random	25	23	48
Total	50	47	97

The significant amount of practice main effect was followed by pairwise comparisons using Tukey's HSD. The largest difference was between the low and medium amount conditions: the medium amount condition had lower RMSE than the low amount condition, but this was not significant using the Tukey HSD adjusted p-value (mean difference = 52.09, $p = .05$). There was not a significant difference between the low and high amount conditions (mean difference = 28.67, $p = .39$), nor was there a significant difference between the medium and high amount conditions (mean difference = 23.43, $p = .53$).

In the random retention block, there was a significant main effect of acquisition schedule, $F_{(1,89)} = 18.10, p < .001, MSE = 7308.78$; on average the random schedule condition had 73 ms lower RMSE than the blocked schedule condition (see Table 18). There was no significant difference between the amount of practice conditions, $F_{(2,89)} = 0.88, p = .42$, nor was there a significant difference between retention order, $F_{(1,89)} = 1.53, p = .22$. There was no significant interaction between practice schedule and retention order, $F_{(1,89)} = 0.28, p = .60$. There was an additional trend in that the difference in RMSE between blocked and random schedules was more prominent in the medium and high amount conditions (mean difference = 89.02 and mean difference = 110.92, respectively) than the low amount condition (mean difference = 20.74). However, this acquisition schedule by amount of practice interaction was not significant, $F_{(2,89)} = 2.56, p = .08$.

3.2.2.2 Pattern Recall

As in Experiment 1, participants completed the paper and pencil recall measure twice: at the end of Session 1 and the beginning of Session 2. A repeated measures analysis was conducted using session as a repeated-measure and acquisition schedule and amount of practice as between-subject factors. There was a significant multivariate

Table 17: Experiment 2. Mean RMSE (and standard deviation) for the blocked retention block, as a function of acquisition schedule and amount of practice.

	Retention order		
Acquisition schedule	Blocked first	Random first	Average
Low Amount			
Blocked	288.73 (64.80)	292.19 (79.07)	290.57 (70.21)
Random	297.42 (63.10)	234.70 (143.97)	267.90 (110.00)
Average	293.61 (61.83)	263.44 (116.07)	278.53 (92.75)
Medium Amount			
Blocked	252.78 (88.35)	221.70 (79.01)	239.18 (83.15)
Random	208.13 (71.78)	218.22 (95.18)	212.84 (80.54)
Average	231.77 (81.76)	219.96 (84.06)	226.43 (81.63)
High Amount			
Blocked	280.20 (85.18)	287.39 (76.53)	283.79 (78.64)
Random	201.55 (92.29)	221.82 (83.94)	211.69 (85.86)
Average	243.19 (94.82)	256.54 (84.55)	249.86 (88.72)
Average			
Blocked	272.71 (79.47)	269.83 (81.04)	271.30 (79.41)$_a$
Random	238.17 (85.90)	225.20 (106.58)	231.96 (95.54)$_b$
Average	255.44 (83.73)	247.99 (96.05)	251.83 (89.52)

Note. Means in each column that share a subscript do not differ significantly ($p < .05$).

Table 18: Experiment 2. Mean RMSE (and standard deviation) for the random retention block, as a function of acquisition schedule and amount of practice.

	Retention order		
Acquisition schedule	Blocked first	Random first	Average
Low Amount			
Blocked	257.23 (70.88)	295.54 (98.78)	277.66 (86.16)
Random	256.14 (65.44)	257.81 (115.67)	256.92 (89.42)
Average	256.61 (65.53)	276.68 (105.72)	266.64 (87.12)
Medium Amount			
Blocked	279.58 (107.23)	288.92 (102.45)	283.67 (101.75)
Random	188.72 (87.50)	201.42 (76.56)	194.64 (79.90)
Average	236.82 (106.23)	245.17 (98.03)	240.59 (101.00)
High Amount			
Blocked	291.92 (55.01)	334.82 (87.88)	313.37 (74.47)
Random	190.87 (51.44)	214.05 (101.59)	202.46 (78.70)
Average	244.37 (73.31)	277.99 (110.62)	261.18 (93.97)
Average			
Blocked	277.77 (79.35)	308.34 (94.02)	292.74 $(87.31)_a$
Random	213.68 (74.09)	225.43 (98.75)	219.31 $(86.04)_b$
Average	245.72 (82.59)	267.77 (104.11)	256.40 (93.80)

Note. Means in each column that share a subscript do not differ significantly ($p < .05$).

effect of session, $V = 0.18, F_{(1,91)} = 20.57, p < .001$. There was no significant multivariate interaction between session and schedule, $V = 0.02, F_{(1,91)} = 2.09, p = .15$, between session and amount of practice, $V = 0.03, F_{(2,91)} = 0.03, p = .22$. Additionally, there was no significant higher-order interaction between session, acquisition schedule, and amount of practice, $V < 0.01, F_{(2,91)} = 0.41, p = 0.66$.

The significant session effect was followed with univariate ANOVAs on each session. In session 1, there was a significant effect of acquisition schedule, $F_{(1,91)} = 39.17, p < .001, MSE = 0.06$; the random acquisition schedule condition had higher recall than the blocked acquisition schedule condition (see Table 19). Additionally, there was a significant effect of amount of practice, $F_{(2,91)} = 13.69, p < .001$. Both the medium and high amount of practice conditions had higher recall than the low amount of practice condition, (mean difference = 0.19, $p = .03$ and mean difference = 0.29, $p < .001$, respectively). There was no significant difference in recall between the high and medium amount of practice condition (mean difference = 0.11, $p = .30$). Across both sessions acquisition schedule and amount of practice affected recall, with the random schedule condition performing better than the blocked schedule condition, and the medium and high amount conditions performing better than the low amount condition.

Session 2 showed the same pattern of results as Session 1. There was a significant effect of acquisition schedule, $F_{(1,91)} = 22.31, p < .001$; the random acquisition schedule condition had higher recall than the blocked acquisition schedule condition (see Table 19). Additionally, there was a significant effect of amount of practice, $F_{(1,91)} = 7.45, p = .001$. Both the medium and high amount of practice conditions had higher recall than the low amount of practice condition (mean difference = 0.18, $p = .039$ and mean difference = 0.22, $p = .007$, respectively). There was no significant difference in recall between the high and medium amount of practice condition (mean difference = 0.04, $p = .85$).

Table 19: Experiment 2. Mean proportional recall accuracy (and standard deviation) as a function of acquisition schedule, session, and amount of practice.

	Session	
Acquisition schedule	1	2
Low Amount		
Blocked	0.12 (0.16)	0.08 (0.12)
Random	0.40 (0.31)	0.31 (0.30)
Med Amount		
Blocked	0.31 (0.19)	0.30 (0.23)
Random	0.60 (0.25)	0.48 (0.26)
High Amount		
Blocked	0.39 (0.29)	0.26 (0.30)
Random	0.75 (0.22)	0.61 (0.32)
Average		
Blocked	0.28 $(0.25)_a$	0.22 $(0.25)_a$
Random	0.58 $(0.30)_b$	0.46 $(0.31)_b$

3.2.3 Double Transfer Task

Four participants were excluded because of software errors (N = 3) or because the participant did not finish (N = 1).

3.2.3.1 Errors

Errors were analyzed using a three-way repeated measures analysis with acquisition schedule and amount of practice as between-subject factors and block as a within-subject factor. Mean errors (and standard deviations) for the acquisition schedule X amount of practice X transfer block cells are displayed in Table 20.

The three-way interaction between acquisition schedule, amount of practice, and block was not significant, $V < 0.01$, $F_{(2,87)} = 0.16$, $p = .85$. There was, however, a significant two-way interaction between acquisition schedule and block, $V = 0.14$, $F_{(1,87)} = 13.58$, $p < .001$. There were no other significant higher order interactions: acquisition schedule and amount of practice, $V < 0.01$, $F_{(2,87)} = 0.21$, $p = .81$; amount of practice and block, $V = 0.03$, $F_{(2,87)} = 1.40$, $p = .25$.

Table 20: Experiment 2. Mean errors (and standard deviation) on the double pattern task as a function of acquisition schedule, block, and amount of practice.

Acquisition Schedule	Block 1	Block 2
Low Amount		
Blocked	12.69 (7.08)	2.46 (2.79)
Random	7.35 (6.89)	2.76 (2.44)
Medium Amount		
Blocked	11.64 (8.20)	4.0 (5.19)
Random	5.27 (4.42)	2.4 (2.10)
High Amount		
Blocked	10.39 (7.98)	3.50 (3.45)
Random	5.25 (5.65)	2.25 (2.08)
Average		
Blocked	11.44 $(7.68)_a$	3.36 (3.88)
Random	6.00 $(5.76)_b$	2.48 (2.18)

Note. Means in each column that share a subscript do not differ significantly ($p < .05$).

The two-way interaction between acquisition schedule and block was followed with separate univariate ANOVAs for each block. In Block 1 there was a significant acquisition schedule main effect, $F_{(1,87)} = 15.43$, $p < .001$, $MSE = 4068.10$, in which the random schedule conditions had lower errors than the blocked schedule conditions (see Table 20). There was no significant effect of amount of practice $F_{(2,87)} = 0.88$, $p = .42$, nor was there a significant interaction between acquisition schedule and amount of practice, $F_{(2,87)} = 0.07$, $p = .93$.

In Block 2 there was no significant acquisition schedule main effect, $F_{(1,87)} = 1.73$, $p = .19$, $MSE = 867.39$, no significant amount of practice main effect, $F_{(2,87)} = 0.18$, $p = .83$, and no significant interaction between acquisition schedule and amount of practice $F_{(2,87)} = 0.76$, $p = .47$.

3.2.3.2 Response Time

As in Experiment 1, response time was analyzed as a proxy for planning time. Response times that were greater than six standard deviations from the mean were trimmed, eliminating five response times (5 of 2232 response times, 0.22% of the data). This trimming procedure removed the most extreme outliers. Participants' median response times were calculated for each transfer block. These median response times were log-transformed prior to analysis to normalize the data by reducing the positive skew in median response times.

Response time was analyzed using a three-way repeated measures analysis with acquisition schedule and amount of practice as between subject factors and block as a within subject factor. Means and standard deviations for the median response times are displayed in Table 21.

The only significant effect was a significant block main effect, $V = 0.06$, $F_{(1,87)} = 5.35$, $p < .05$, with response times decreasing from Block 1 to Block 2 (see Table 21). There was no significant main effect of acquisition schedule $V < 0.01, F_{(1,87)} = 0.09$, $p = .77$ and no significant main effect of amount of practice $V < 0.01$, $F_{(2,87)} = 0.66$, $p = .52$. Furthermore, there were no significant interactions between acquisition schedule and amount of practice, $V = 0.03$, $F_{(2,87)} = 1.45$, $p = .24$, between acquisition schedule and block, $V = 0.03$, $F_{(1,87)} = 2.64$, $p = .11$, nor between amount of practice and block, $V = 0.01$, $F_{(2,87)} = 0.65$, $p = .53$. Finally, there was no significant interaction between acquisition schedule, amount of practice, and block, $V = 0.01$, $F_{(2,87)} = 0.50$, $p = .61$.

3.2.4 Mirror Transfer Task

Two participants were excluded because of software errors (N = 1) and because the participant did not finish the task (N = 1).

Movement times that were more than four standard deviations away from the

Table 21: Experiment 2. Mean median response time (and standard deviation) on the double pattern task as a function of acquisition schedule, block, and amount of practice.

Acquisition Schedule	Block 1	Block 2
Low Amount		
Blocked	4998.73 (4832.89)	4317.08 (1905.92)
Random	5969.15 (4098.36)	4657.91 (3761.74)
Medium Amount		
Blocked	3568.29 (1624.54)	3227.89 (1250.84)
Random	4604.43 (1526.45)	3748.17 (1593.75)
High Amount		
Blocked	4509.75 (2923.73)	3995.89 (1627.38)
Random	3401.84 (1496.26)	3214.56 (1533.35)
Average		
Blocked	4358.11 (3283.54)	3849.74 (1634.78)
Random	4686.91 (2878.47)	3892.50 (2588.29)

mean (for each task) were trimmed, reducing positive skew in data. Data from 17 trials were trimmed (17 of 2280 total trials were trimmed, 0.75% of the mirror transfer data). RMSE was calculated for each block. RMSE was analyzed using three-way repeated measures analysis with acquisition schedule and amount of practice as between-subject factors and block as a within subject factor. RMSE descriptive statistics are shown in Table 22.

There was a significant block effect, $V = 0.26$, $F_{(1,89)} = 32.94$, $p < .001$, with RMSE decreasing from Block 1 to Block 2. Additionally, there was a significant acquisition schedule effect, $V = 0.08$, $F_{(1,89)} = 7.47$, $p < .01$; the random schedule condition had lower RMSE than the blocked schedule condition.

There was no significant amount of practice main effect, $V = 0.15$, $F_{(2,89)} = 0.68$, $p = .51$. There was no significant interaction between acquisition schedule and amount of practice, $V < 0.01$, $F_{(2,89)} = 0.22$, $p = .80$, nor between acquisition schedule and block, $V = 0.39$, $F_{(1,89)} = 0.39$, $p = .53$. Finally, there was no significant

Table 22: Experiment 2. Mean RMSE (and standard deviation) on the mirror transfer task, as a function of acquisition schedule and block.

Acquisition schedule	Block 1	Block 2
Low Amount		
Blocked	290.69 (85.15)	206.54 (86.55)
Random	240.72 (101.26)	187.07 (90.26)
Medium Amount		
Blocked	303.62 (126.99)	258.16 (142.64)
Random	249.42 (104.39)	189.12 (100.95)
High Amount		
Blocked	272.63 (84.55)	234.85 (89.91)
Random	201.88 (124.13)	180.25 (96.85)
Average		
Blocked	287.90 (98.86)	233.85 (108.24)
Random	230.49 (109.93)	185.44 (93.90)

interaction between acquisition schedule, amount of practice, and block, $V < 0.01$, $F_{(2,89)} = 0.55$, $p = .57$.

3.3 Discussion

During acquisition the blocked practice schedule decreases RMSE. Additionally, medium and high amounts of practice result in lower RMSE, relative to low amounts of practice. These data indicate that, for all the amounts of practice used in this experiment, the blocked schedule improves acquisition performance. Additionally, increasing practice from the low amount to either the medium or high amount of practice results in continued performance improvement for both the blocked and random acquisition schedules.

The retention measures show that practice schedule affects RMSE. As in Experiment 1, the random acquisition schedule results in lower RMSE during the random retention block, relative to the blocked acquisition schedule. Unlike in Experiment 1, the random acquisition schedule condition also results in lower RMSE during the blocked retention block, relative to the blocked acquisition schedule (i.e., for the measure reflecting both retention and transfer of practice schedule).

The pattern recall data show the same amount of practice trends as acquisition performance. Extending practice from a low amount to a medium or high amount improves performance on the immediate and delayed recall test. Additionally, the recall data show the opposite effect of acquisition performance: the random schedule condition has higher recall than the blocked schedule condition.

The transfer measures also show practice schedule effects but no amount of practice effects. On the double transfer task, practice schedule affects the number of errors (incorrect keypresses). In Block 1, the blocked acquisition schedule condition has almost twice as many errors as the random acquisition schedule condition. This is consistent with the pattern recall data, suggesting one source of incorrect keypresses is difficulty in using the goal movement time stimuli to retrieve the appropriate keypress sequence. Another potential source of errors is a difficulty in creating and programming the novel action plan to sequence the 10 keys. On the mirror transfer task, practice schedule affects RMSE. The random acquisition schedule condition is better able to produce a mirror-reversed keypress sequence according to its goal time, perhaps because of improved facility at retrieving or manipulating motor programs in working memory, reducing the need for dynamic motor programming during response execution.

Across all the measures there was no interaction between practice schedule and amount of practice. There were, however, trends towards an interaction in a couple of measures. Notably, there was a trend in the measure that is most likely to show these effects: performance on retention blocks ordered according to a random practice schedule. On this measure the mean difference between the blocked and random schedules was about 20 ms for the low amount condition, but approximately 100 ms for the medium and high amount conditions. This suggests that at low amounts of practice the retention benefits afforded by a random acquisition schedule might not yet appear.

If one were to have conducted a study using this task, but only a low amount of practice, the practice schedule effects seen in the RMSE retention measures (both blocked and random retention blocks) would not have been significant. The acquisition and retention measures show that performance is still improving from the low to medium and high amounts of practice, and the retention measures show the benefits do not reliably appear until longer training durations. This suggests that a minimum amount of training is needed for practice schedule effects to appear. Extending this idea, these results are consistent with the claim that one reason complex motor skill acquisition studies often do not show practice schedule effects is that thy might not have had long enough acquisition training to see the effects appear in retention measures.

CHAPTER IV

EXPERIMENT 3

4.1 Introduction

Experiment 3 used a novel perceptual categorization task in which participants learned to categorize football defensive play formation diagrams.

Experiment 3 was conducted to determine (1) the number of practice trials for each practice amount condition in Experiment 4 and (2) which form of a blocked practice schedule (i.e., a blocked or a blocked-repeated) to use in Experiment 4.

4.2 Method

4.2.1 Participants

Forty-seven participants over the age of 17 were recruited from the Georgia Tech participant pool with the exclusion criteria that they had not "played on a football team in high school or college".

Five of these participants were excluded prior to data analysis for a variety of reasons: did not finish the Session 1 training (N = 3), software crashed during training (N = 1), and wrote notes to himself about play category rules (N = 1). Three additional participants were excluded because their data suggested they were "gaming the system" by ignoring task instructions in order to more quickly finish the experiment. All three participants were in the random condition. Appendix C explains the system and the criteria used to separate these participants.

Excluding these participants resulted in 39 participants; cell sizes are displayed in Table 23. Participants' mean age was 20.0 years ($SD = 1.5$). Of the 39 participants, 15 were male, 18 were female, and 6 did not answer.

Table 23: Experiment 3. Number of participants (during Session 1, acquisition) in each acquisition schedule and acquisition schedule version cell.

	Version				
Acquisition schedule	1	2	3	4	Total
Blocked	3	3	4	3	13
Blocked-repeated	4	4	4	3	15
Random	6	5	–	–	11

Note. The random schedule condition had only two versions.

4.2.2 Materials

4.2.2.1 Task

On each trial the software displayed a play label (i.e., its category name) and a play diagram (an example instance of the category; see Figure 17). A diagram that matches the play label is a *positive example* of the category and a diagram that does not match the category is a *negative example* (Bourne, 1970). The participant's task was to decide whether the play diagram was a positive or negative example instance of the play category by classifying the diagram as "legitimate" or "not legitimate" (respectively). After responding, the software displayed feedback indicating whether the participant's response was accurate.

Prior to beginning the task, participants were told which attributes are relevant (i.e., number of squares and position of circles) but not what rules guide feature value assignments to category membership (e.g., all plays in the "Blue" category have three squares and one circle in the backfield). The task is thus an induction (rule learning) task in which participants are given the set of features and learn the rules that determine category membership. Participants use this conceptual structure to classify subsequent example instances.

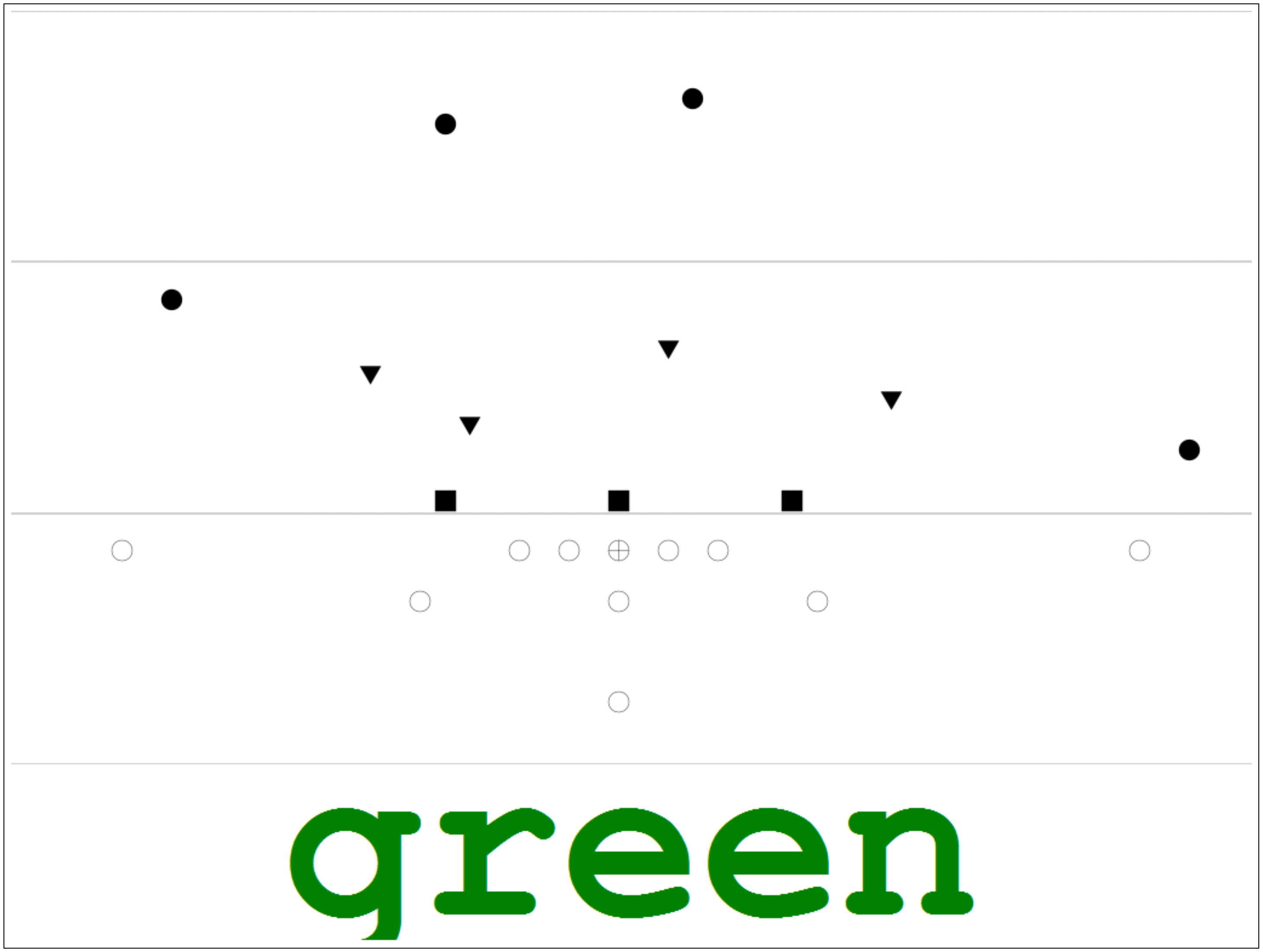

Figure 17: Sample stimulus for a green play category (positive example instance) in Experiments 3 and 4.

Table 24: Experiment 3. The four acquisition play categories, with relevant and irrelevant features.

		Relevant		Irrelevant		
Phase	Label	DL	Safety in backfield	Odd	SS	Canonical name
Acq.	Blue	3	1	(**Y**, N)	(L, R)	3–4 Cover 3
Acq.	Gray	4	2	(Y, **N**)	—	4–3 Cover 2
Acq.	Green	3	2	(**Y**, N)	—	3–4 Cover 2
Acq.	Red	4	1	(Y, **N**)	(L, R)	4–3 Cover 3

Note. Relevant feature values determine category membership. Irrelevant feature values are randomly selected for each example instance. If a set of feature values has a bold value, this value has a higher probability of being selected when creating example instances.

4.2.2.2 Stimuli

Four categories of defensive plays (front formation + secondary coverage) were used in the acquisition trials. (Figures 18, 19, 20, and 21) show a prototype diagram for the blue, gray, green, and red categories, respectively. In these plays, category membership is determined by the cooccurance of two bidimensional features: number of down linemen and number of safeties in the backfield (see Table 24; also, Table 25 lists defensive player positions, their number, and abbreviations). Two *relevant features* are sufficient to categorize the play (see Table 24). In addition, there are two *irrelevant features* that do not affect category membership; these irrelevant features are also bidimensional (see Table 24). The combination of irrelevant features allows a category's example instances to vary at a gross-level. The number of relevant and irrelevant dimensions is similar to previous concept learning research (e.g., Kurtz & Hovland, 1956).

Creating example instances Each play category has multiple example instances (i.e., diagrams) that are created by (1) varying the irrelevant features, and (2) varying the position of the players. Position of the players is randomly distorted, adapting the established procedures for distorting random dot patterns (Posner, Goldsmith,

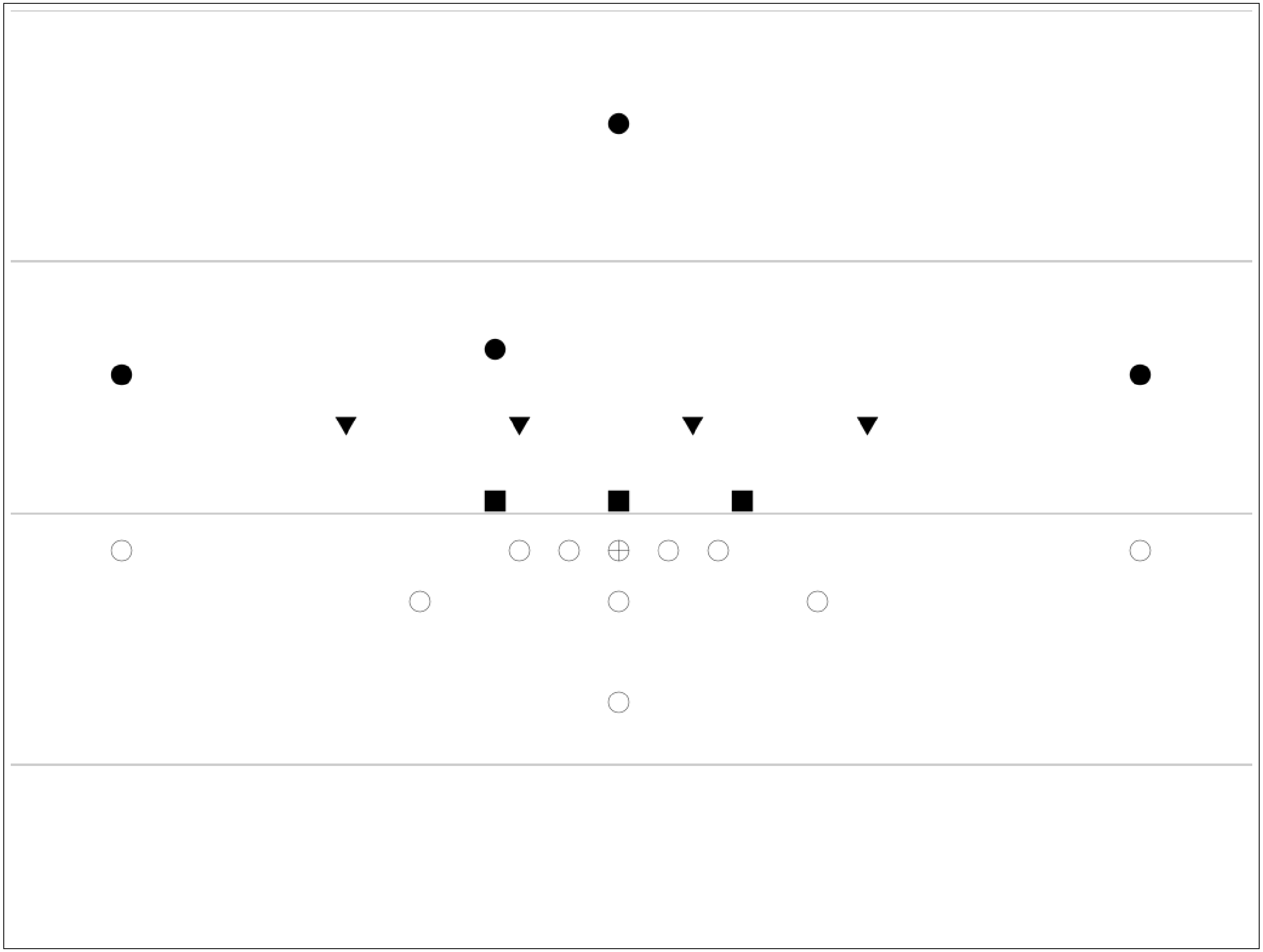

Figure 18: Prototype play diagram for the blue play category. The blue play category has four down lineman (squares) and one safety (a FS) in the backfield.

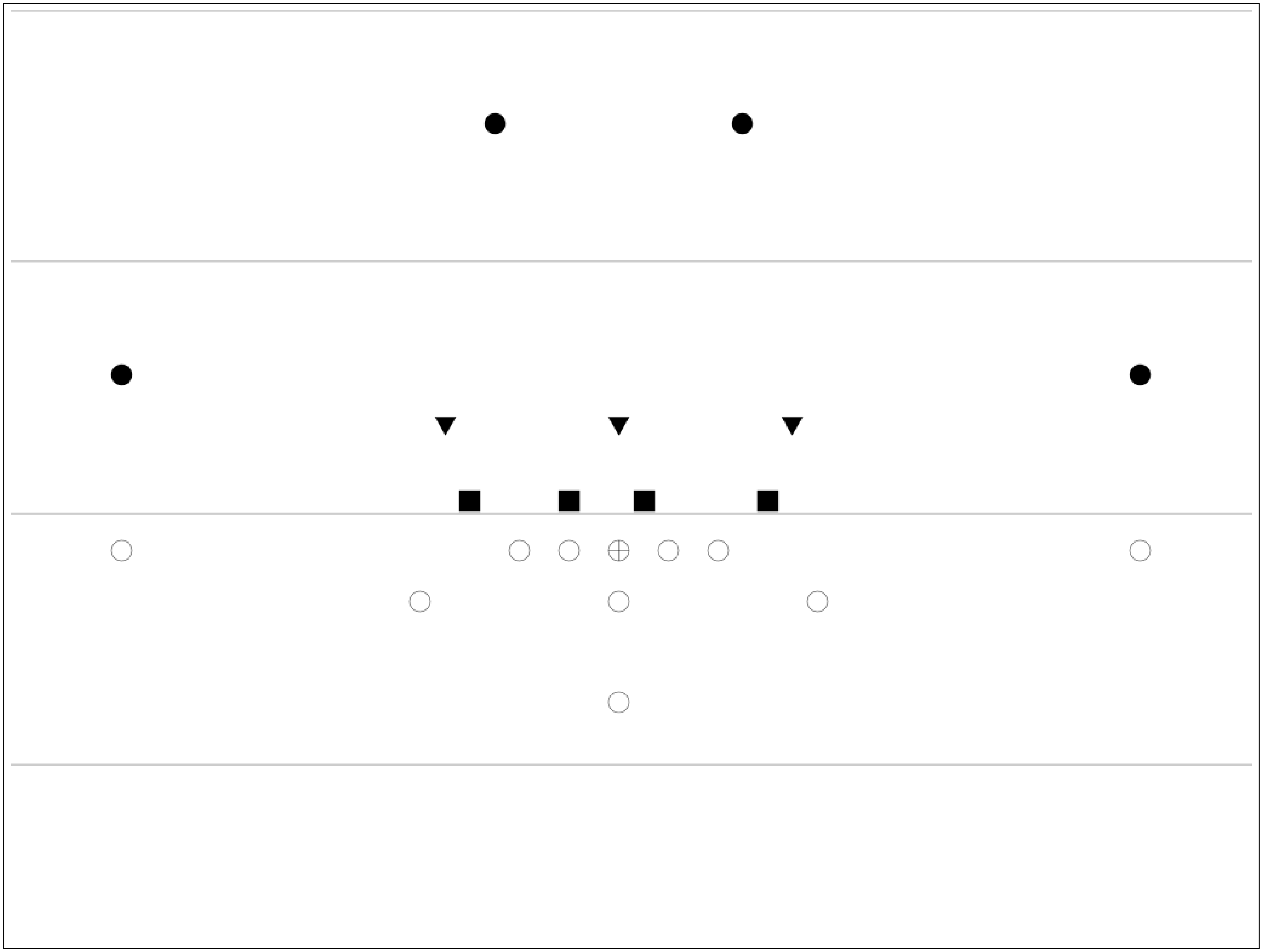

Figure 19: Prototype play diagram for the gray play category. The gray play category has four down lineman (squares) and two safeties in the backfield.

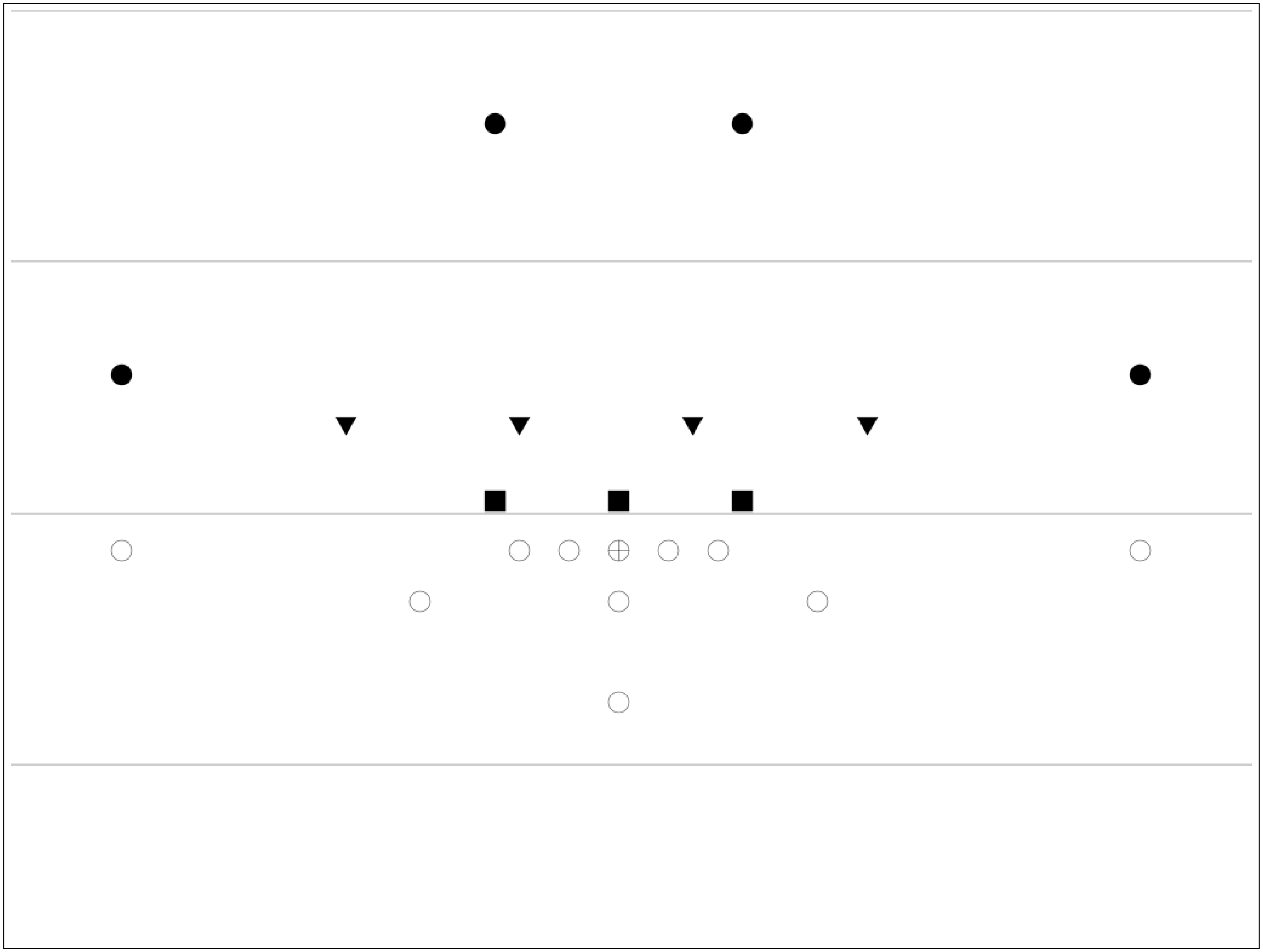

Figure 20: Prototype play diagram for the green play category. The green play category has three down lineman (squares) and two safeties in the backfield.

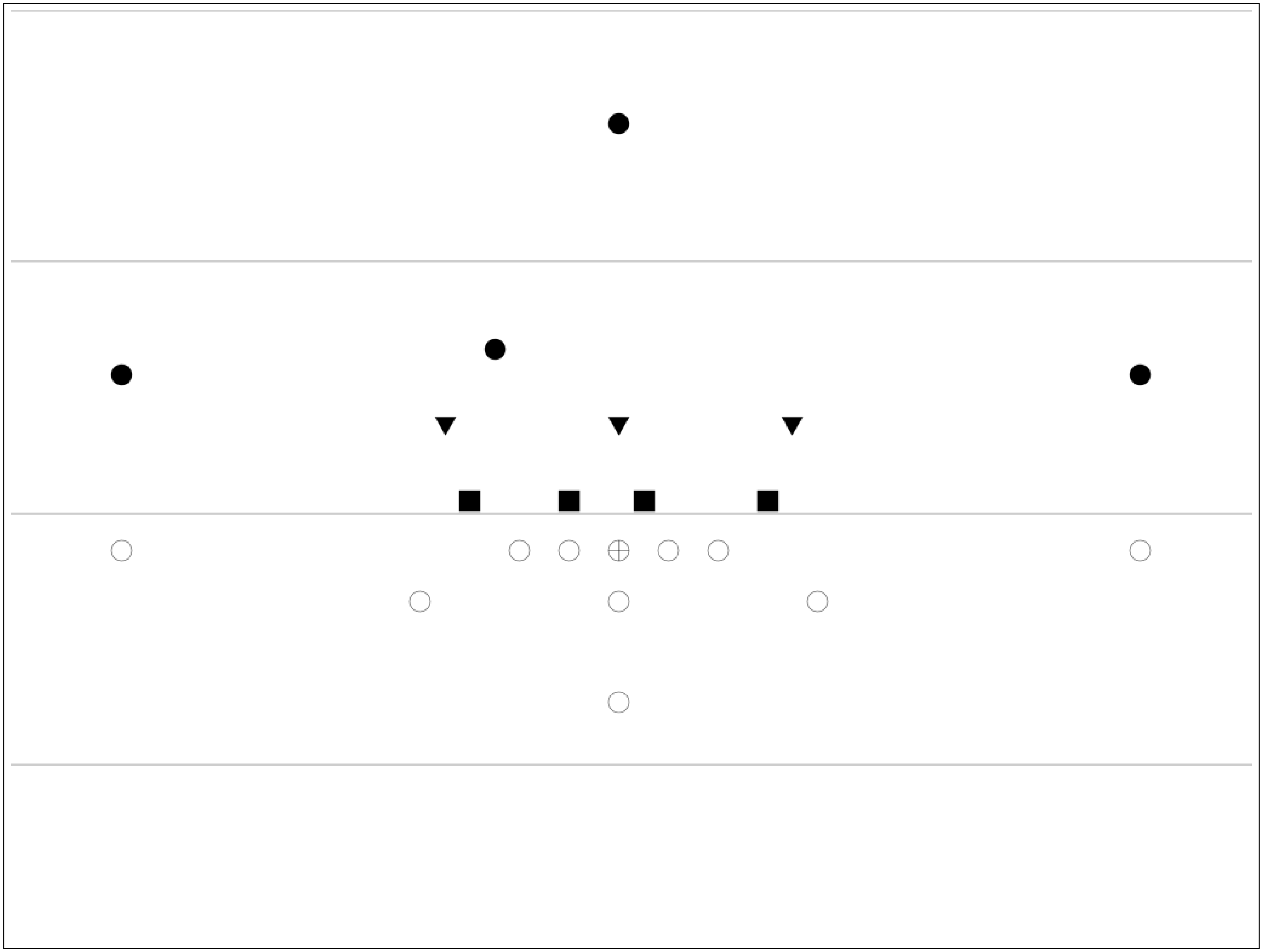

Figure 21: Prototype play diagram for the red play category. The red play category has four down lineman (squares) and one safety (a FS) in the backfield.

Table 25: Abbreviations and number of (possible) players for each defensive position.

Position	Abbreviation	Number
Down lineman	DL	(3–5)
Linebacker	LB	(2–4)
Cornerback	CB	2
Free safety	FS	1
Strong safety	SS	1
Nickel back	NB	(0–1)

Note. Play categories differ in the number of player positions they use. Therefore some positions have a range of possible players (in parenthesis). Free safety and strong safety are both subdivisions of the broader category, Safety (S).

& Welton, 1967; J. D. Smith, Redford, Gent, & Washburn, 2005). In this study, the distortion procedure was modified to allow constraints of realistic football player movements (e.g., a down lineman may move horizontally along the line of scrimmage but may not move in front of the line of scrimmage, players may not overlap one another, etc.). Additionally, the variability of individual player movements is dependent on the player position. For example, a cornerback has 35 possible locations (a 7 x 5 rectangle) whereas a safety has 81 possible locations (a 9 x 9 square). Overall, the conceptual structure and procedure for creating example instances adapts methods from the concept learning and perceptual learning literature.

Each example instance is created by setting initial player locations based on the play category prototype. Next, gross-level player positions are modified by randomly sampling irrelevant features from probability distributions of irrelevant feature values. Finally, fine-level player positions are randomly distorted (within their regions of distortion) by applying the distortion algorithm. Figures 22 and 23 show two example instances of the green play category and Figures 24 and 25 show two example instances of the red play category.

Importantly, both positive and negative instances are created in reference to the

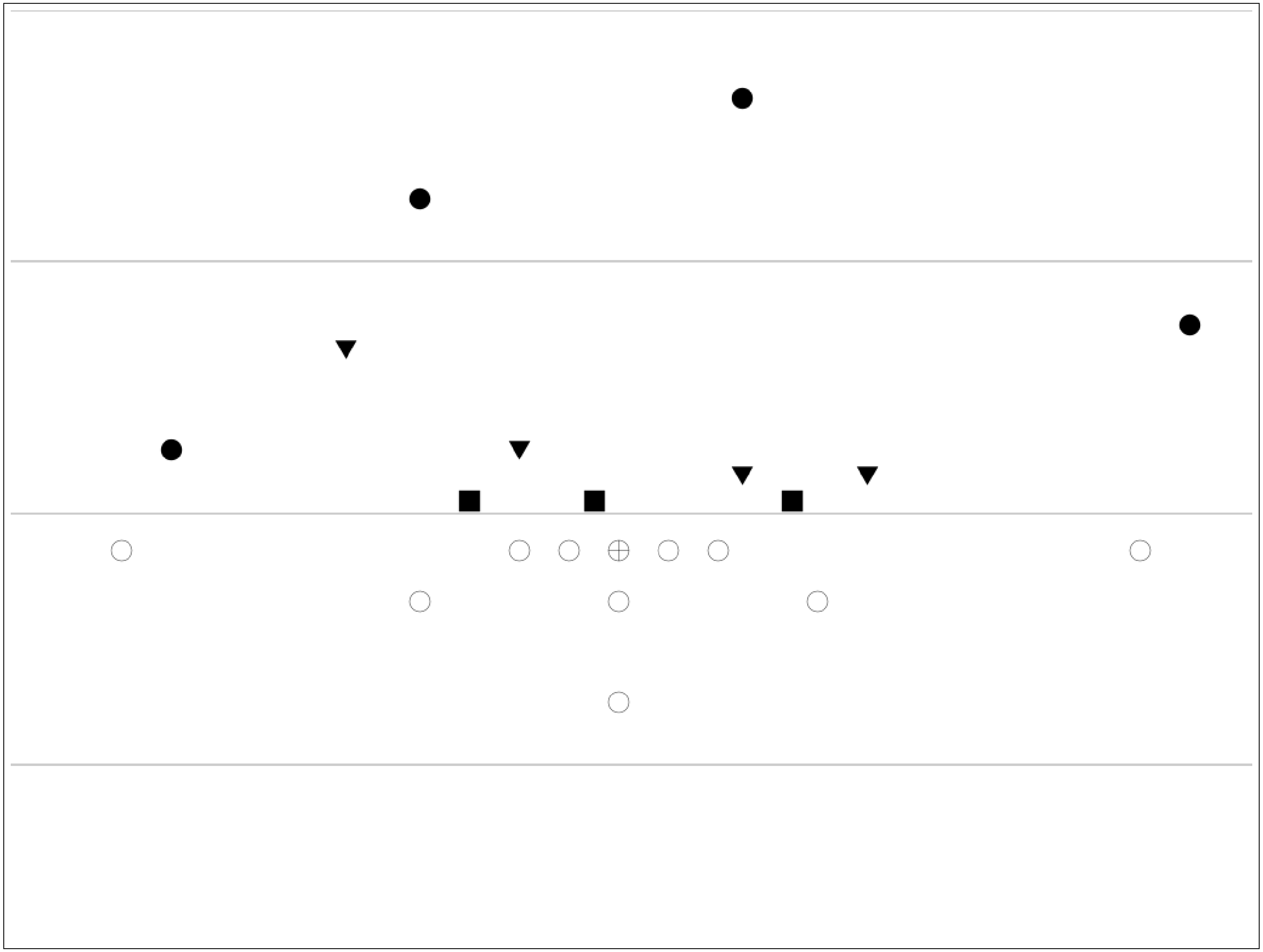

Figure 22: Example instance for the green play category. In this example, there are three down lineman (DLs) and two safeties (Ss) in the backfield (both relevant features). Additionally, the nose tackle (a DL) is aligned in an even set.

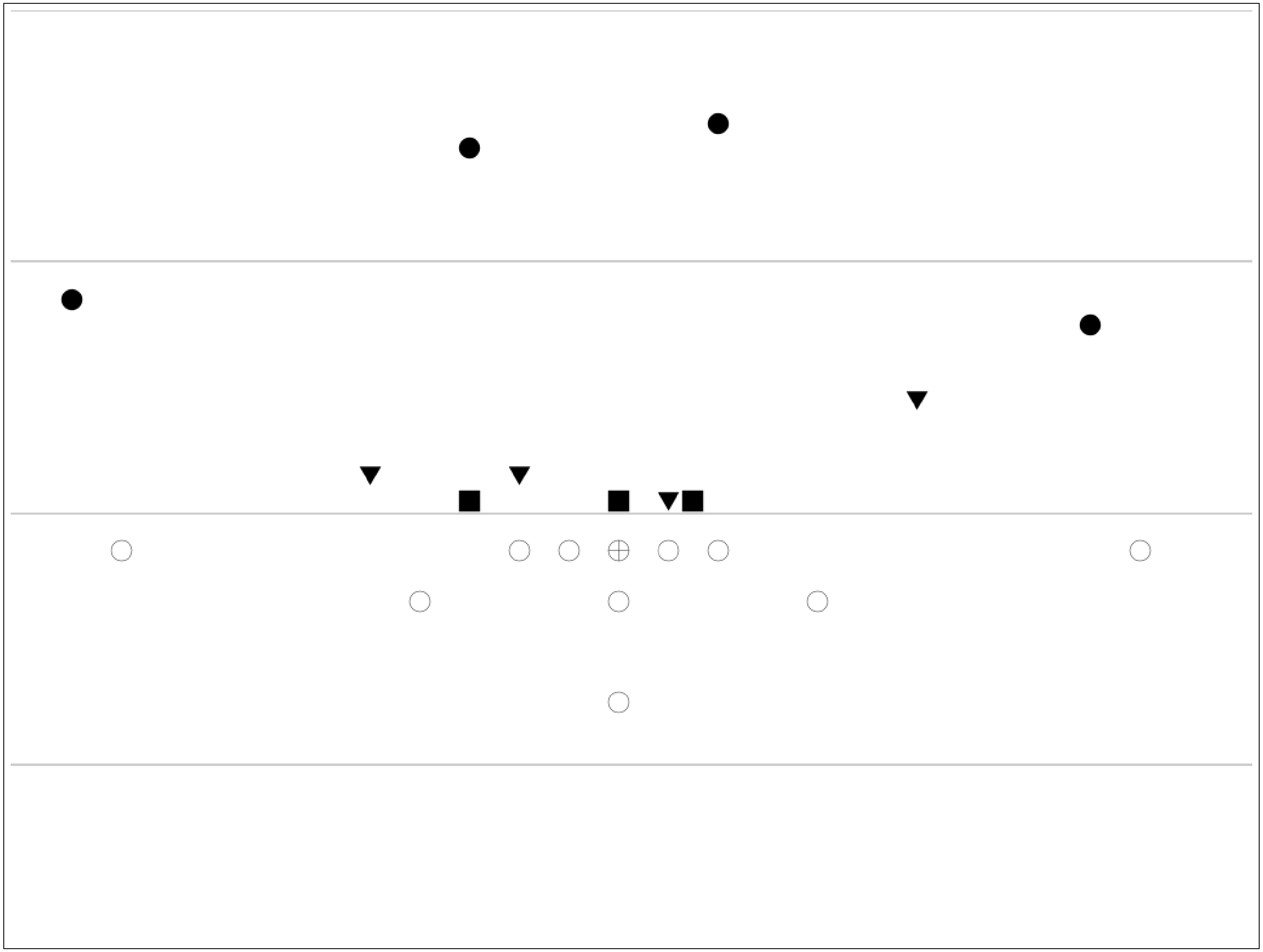

Figure 23: Example instance for the green play category. In this example, there are three down lineman (DLs) and two safeties (Ss) in the backfield (both relevant features). Additionally, the nose tackle (a DL) is aligned in an odd set.

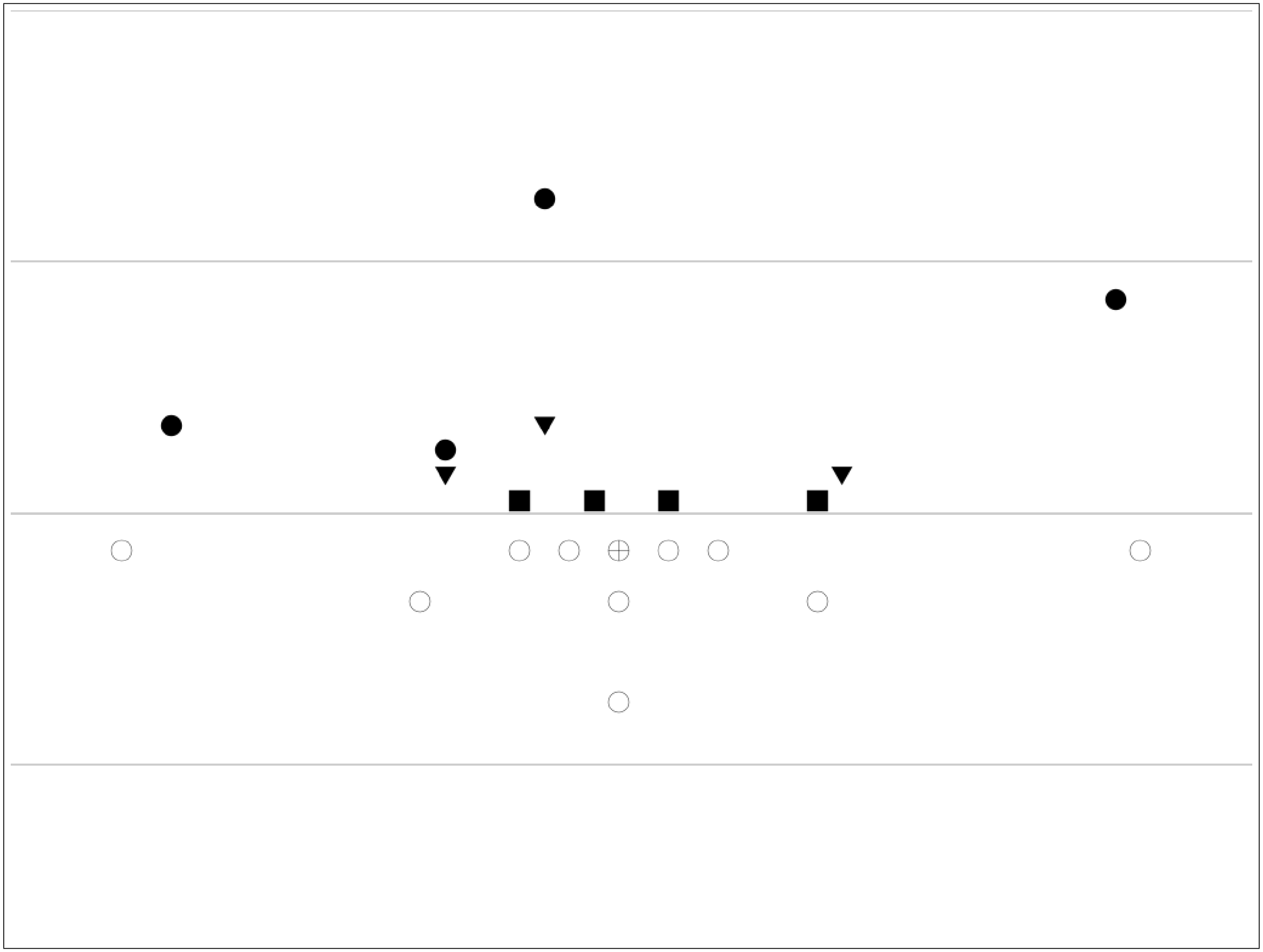

Figure 24: Example instance for the red play category. In this example, there are four down lineman (DLs) and one safety (FS) in the backfield (both relevant features). Additionally, the nose tackle (a DL) is aligned in an even set and the strong safety (SS) is aligned to the left of the center (both irrelevant features).

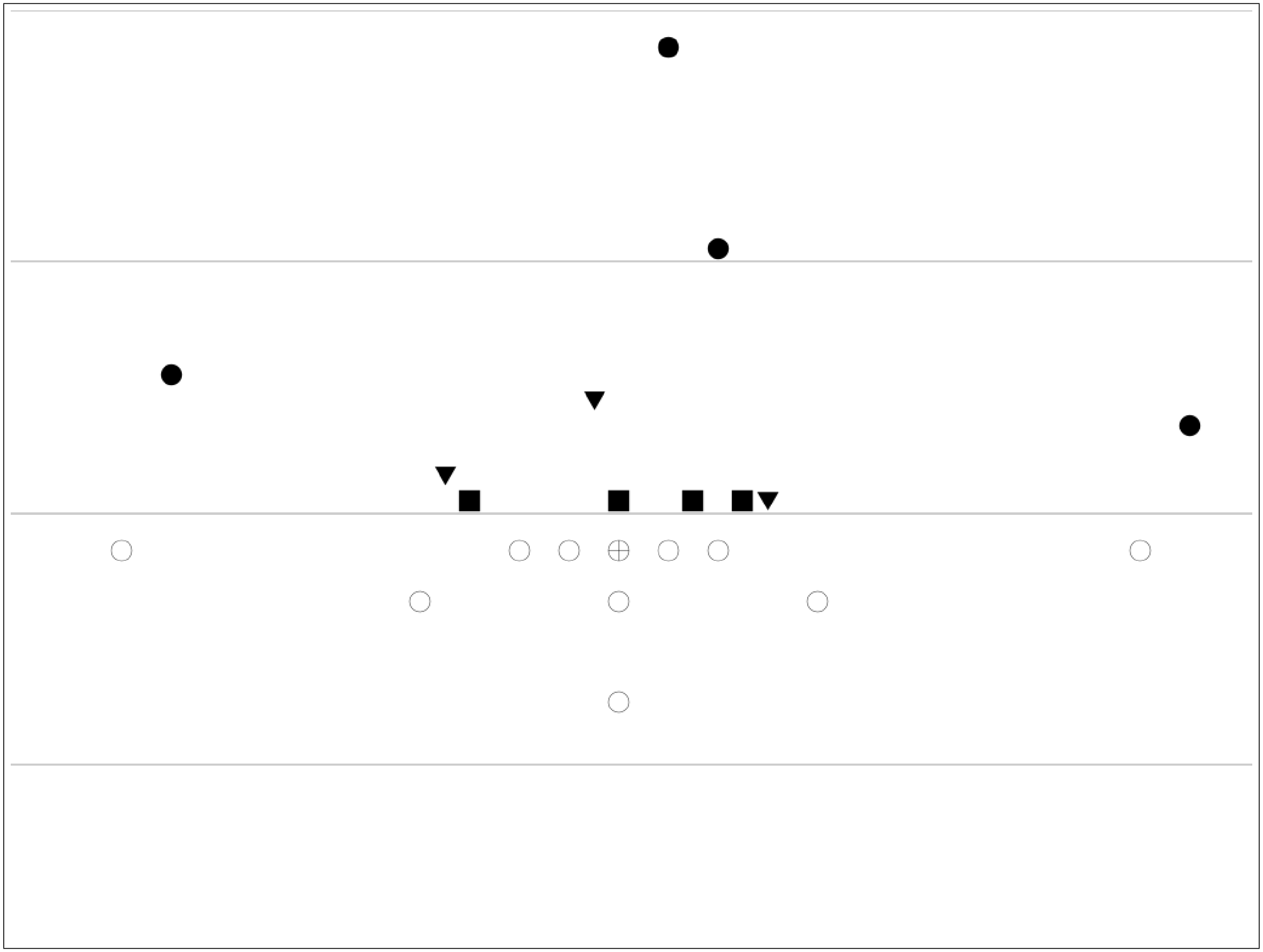

Figure 25: Example instance for the red play category. In this example, there are four down lineman (DLs) and one safety (FS) in the backfield (both relevant features). Additionally, the nose tackle (a DL) is aligned in an odd set and the strong safety (SS) is aligned to the right of the center (both irrelevant features).

play category. That is, the play label always matches the play category. Thus, any time the play diagram does not match the label, the diagram is a negative example instance of the category. For example, if the play category is black then the play label is always black. If the play diagram (e.g., a distortion of the black category) matches the play label, then the diagram is a positive example. If the play diagram (e.g., a distortion of the green category) does not match the play label, then the diagram is a negative example.

4.2.2.3 Questionnaires

A football knowledge questionnaire (see Appendix D) was used to determine participants' familiarity with football.

4.2.3 Design

4.2.3.1 Practice schedule

Three practice schedules were used: blocked, blocked-repeated, and random. In both the blocked and blocked-repeated schedules, each training block used one play category and participants classified diagrams as positive or negative examples of the category. In the random schedule, each training block used all four play categories and participants classified diagrams as positive or negative examples of each category.

As in Experiment 1, multiple versions of the acquisition schedules were used. Four blocked, four blocked-repeated, and two random schedules were used. The blocked and blocked-repeated schedules used four versions (instead of using only two, as in Experiment 1) because of the potential that order effects might arise from the specific sequence of play categories *within* a schedule version. Recall that two valid features (number of down lineman/squares and position of safeties/circles) may vary between play categories. When sequencing play categories within a blocked schedule, one can order categories such that either the number of down linemen changes and/or the position of the safeties changes between consecutive categories. The specific

Table 26: Experiment 3. Order of play categories in each schedule version, for the blocked and blocked-repeated schedules.

	Category position			
Version	1	2	3	4
1	Green	Blue	Gray	Red
2	Blue	Red	Green	Gray
3	Gray	Green	Red	Blue
4	Red	Gray	Blue	Green

category sequences in each version were selected by (1) using a Latin square to place each category at one of four ordered positions within the schedule versions (i.e., first position, second position, etc.) and (2) selecting a set of sequences such that moving from the first category to the second category changed only one feature, moving from the second category to the third category changed both features, and moving from the third category to the fourth category changed one feature. Table 26 shows the sequence of categories for the four schedule versions, which follow a Latin square. Table 27 shows, for each schedule version, which features change between consecutive categories within a schedule. The blocked-repeated versions used the same category sequences as the blocked versions, but those sequences repeated three times.

Distributing positive and negative example instances within a block In acquisition, each block consisted of 19 trials: 16 positive examples and 3 negative examples. Negative examples were used to ensure that participants (especially in the blocked condition) were actively processing the examples to determine which features determine category membership. If no negative examples were used, participants in the blocked condition might realize that each training block had example instances for only one category, and perform the task without actively processing the examples.

Selecting positive example categories In blocked and blocked-repeated schedules the order of the four categories was assigned to blocks by counterbalancing over

Table 27: Experiment 3. Feature changes that occur after category switches, for each practice schedule version.

	Category 1 vs. Category 2		Category 2 vs. Category 3		Category 3 vs. Category 4	
Version	DL change	S change	DL change	S change	DL change	S change
1	N	Y	Y	Y	N	Y
2	Y	N	Y	Y	Y	N
3	Y	N	Y	Y	Y	N
4	N	Y	Y	Y	N	Y

Note. Each schedule version has one feature change between Category 1 and 2, two feature changes between Category 2 and 3, and one feature change between Category 3 and 4.

participants (see Table 26). In the blocked schedule all blocks for a category were completed before a new category was introduced. In the blocked-repeated schedule one-third of the blocks for a category were completed before a new category was introduced.

In the random schedule, all four play categories were assigned to a block. On each trial the play category was randomly selected with two constraints: (1) a play category could not appear consecutively on more than two trials, and (2) within each block there were an equal number of trials of each category.

Selecting negative example categories In both the blocked and blocked-repeated schedule conditions, negative example trials were created from the same category as the positive instances (i.e., all the positive and negative examples in a block have the same category label).

In the random schedule condition, the negative example trials were also created from the same categories as the positive instances, but there are multiple categories of positive instances. Therefore, distributing these negative example trials within a block followed a more complex algorithm than the blocked schedule. The set of four play categories forms a four element pool. On any given negative example trial, an element from this pool was randomly sampled without replacement with the constraint that no more than two consecutive trials (positive or negative) used the same category. After four negative example trials the pool was repopulated for the next round of sampling; the state of the pool was preserved across blocks. Therefore, on every set of four negative examples, each play category was used once.

Selecting negative example diagrams In the blocked and blocked-repeated schedule conditions the diagrams used to create the negative example instances were selected (without replacement) from a three-category pool. This pool had the three categories that did not match the category of positive example instances for that block.

In the random schedule conditions the diagrams used to create the negative example instances were selected (without replacement) from a four-category pool, with the constraint that the diagram did not match the category of the play label (otherwise it would not be a negative example). The pool was repopulated when depleted and the state of the pool was preserved across blocks.

Appendix E includes more detailed information on the mechanics of distributing examples within each practice schedule and how these distributions were managed across multiple versions within each schedule.

4.2.3.2 Amount of practice

As in Experiment 1, amount of practice (i.e., low, medium, and high) was not manipulated. Instead, participants completed 24 blocks with 19 trials per block, yielding 456 trials. The majority of these trials were positive example instances (n = 384), with a smaller number of negative example instances (n = 72; 16% of all trials).

4.2.3.3 Retention interval

Participants completed two sessions: Session 1 (acquisition), and 48 hours later, Session 2 (retention and transfer).

4.2.4 Procedure

4.2.4.1 Session 1

Participants first completed a demographic questionnaire and the football knowledge questionnaire. The experimenter then explained the general procedure for the first session and the task. In particular, the experimenter explained that participants would be viewing football play diagrams, that the diagrams represent different defensive play formations, and that their goal was to learn the "rules or features that determine what makes a play diagram a legitimate example of each play type." They were also told that on some trials the diagram would not match the label (and should be considered a "not legitimate example"). The experimenter answered any participant questions and then started the experimental software.

The experiment software displayed additional instructions to participants regarding the task and goals. The instructions began by focusing participants on the diagram attributes to attend to during learning: "Each of these four plays has a unique combination of 1) number of squares 2) position of circles." The instructions then showed (1) a sample play with the squares highlighted in red, (2) the same sample play with the circles highlighted, and (3) a split screen showing Images 1 and 2 plus text that repeated that each play was a unique combination of number of squares and position of circles. Additionally, the software instructions repeated what participants' goals and priorities should be when learning the task: "1. To learn the rules or features that determine what makes a diagram a legitimate example of each play type. 2. To improve the accuracy with which you classify each diagram. 3. To increase the speed with which you make your decisions."

The explicit instructions about the play category features and the ranked ordering of goals was added after pilot testing indicated a substantial portion of participants were unable to acquire the rules for each category and that some participants were sacrificing accuracy for speed of response.

A trial began with the display of a fixation point for 2 s. The stimuli then appeared (play diagram and play label) and remained on the screen until the participant responded. Feedback was displayed for 4 s[1]. After a pause the next trial began (intertrial interval = 2.5 s). Incorrect trials were repeated immediately.

Feedback included three pieces of information:

1. Overall accuracy (the total percentage correct for the epoch)

[1]Similar to Experiment 1, feedback duration was decreased to 3 s once the second epoch began.

2. Correct streak (the number of consecutive, correct trials for the epoch)

3. Decision time (the response time, reported in milliseconds)

The design of the feedback was revised after piloting in order to encourage accuracy, rather than speed, of responses.

As in Experiment 1, Session 1 was divided into three epochs. Each epoch had eight blocks. Between epochs participants completed the same paper and pencil distracter tasks (Digit-Symbol Substitution test, Word Beginnings and Endings test, and Making Sentences test) as in Experiment 1[2].

At the end of Session 1 participants completed a paper and pencil recall test in which they had to describe and draw the four play categories.

4.2.4.2 Session 2

After completing Session 1, participants returned 48 hours later to complete Session 2.

When participants returned for Session 2 they completed the paper and pencil recall test again, followed by a retention task and two transfer tasks.

Retention task As with acquisition, retention trials displayed a category label and category diagram; participants classified each diagram as a positive or negative example. Feedback was not given on retention trials and incorrect trials were not repeated. Participants completed two retention blocks with 19 trials per block. The retention blocks used the same categories as acquisition, but different diagrams (i.e., novel example instances). One of the retention blocks was ordered using a blocked schedule and the other retention block was ordered using a random schedule. The order of the blocked and random retention blocks were counterbalanced over participants. As in Experiment 1, the order of categories within the blocked or random

[2]Participants completed only the first page of both the Word Beginning and Endings test and the Making Sentences test.

retention trials was matched to acquisition schedule version. Appendix E explains the retention schedules in more detail.

Category assignment task The first transfer task used a variation of the acquisition task. Instead of categorizing example instances as positive or negative examples of one specific category, participants identified the category of each example instance. In this *category assignment task*, participants saw a play diagram without a corresponding label. Participants categorized the play diagram (e.g., the diagram is an instance of the black play category). The response keys assigned to each category label were displayed on the bottom of the screen on all trials: [F] = Blue, [G] = Gray, [H] = Green, [J] = Red.

Participants completed two blocks of the category assignment task. Both blocks used a random schedule; all participants used this same schedule. Each block had 16 trials, with four example instances for each of the four acquisition play categories. Participants received no feedback after their response and incorrect trials were not repeated.

The category assignment task was designed to provide information regarding the generalizability of the training procedure. It allows one to assess the similarity between classifying examples as belonging to or not belonging to a category (i.e., the acquisition task) and directly assigning examples to a category (i.e., the transfer task). The latter task is more similar to the common usage of the term "categorizing". Because the category names and response key mappings are displayed on the screen, participants do not have to recall category names nor response keys. The additional task demand introduced with this category assignment task is primarily due to the number of response alternatives.

Novel categories task Participants also completed a transfer task in which they learned to classify example instances of three new play categories. The task matched

Table 28: Experiment 3. The two consistent transfer play categories, with relevant and irrelevant features.

		Relevant		Irrelevant		
Phase	Label	DL	Safety in backfield	Odd	SS	Canonical name
Trans.	Purple	5	1	(**Y**, N)	(L, R)	5–2 Cover 3
Trans.	Yellow	5	2	(**Y**, N)	—	5–2 Cover 2

Note. Relevant feature values determine category membership. Irrelevant feature values are randomly selected for each example instance. If a set of feature values has a bold value, this value has a higher probability of being selected when creating example instances.

Table 29: Experiment 3. The inconsistent transfer play category, with relevant and irrelevant features.

		Relevant	Irrelevant		
Phase	Label	NB	DL	Safety in backfield	Canonical name
Trans.	Purple	Yes	(3, 4)	2	Nickel

Note. Relevant feature values determine category membership. Irrelevant feature values are randomly selected for each example instance.

the acquisition task: the transfer play's category label was displayed and participants decided whether the diagram was a positive or negative instance of the category.

Two of these novel play categories are *consistent* with the categorization scheme used for the acquisition categories (see Table 28). One of the novel plays categories is *inconsistent* with the categorization scheme (see Table 29). The consistent plays (purple and yellow) are categorized based on the number of down lineman (5) and the number of safeties in the back field (1 = purple and 2 = yellow). Therefore, the novel, consistent play categories and the acquisition play categories use the same set of relevant features: the number of down lineman (3, 4, or 5) and the number of safeties (1 or 2). Figures 26 and 27 show a prototype of the purple and yellow play categories, respectively.

The inconsistent play, on the other hand, uses a different categorization scheme.

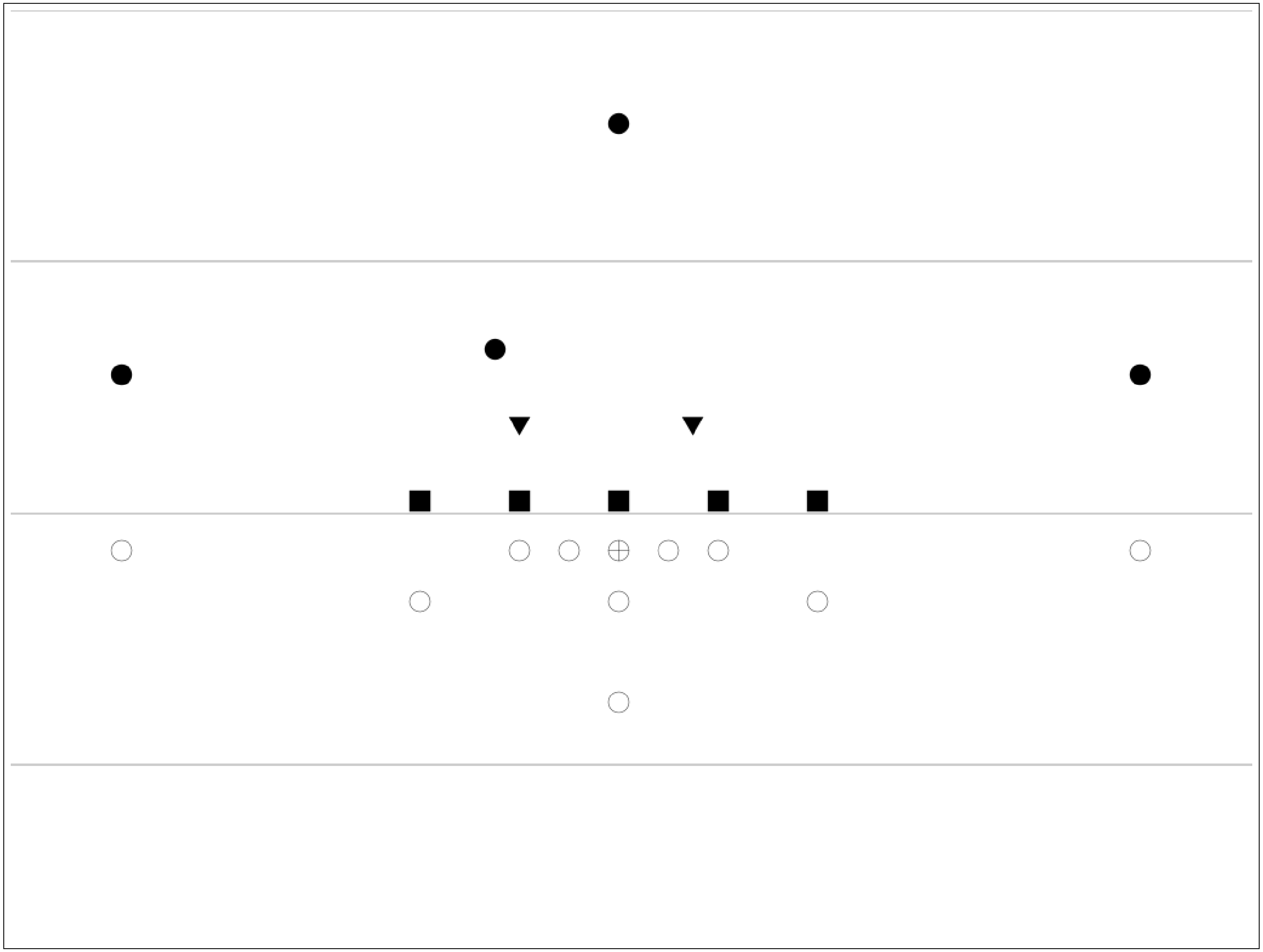

Figure 26: Prototype play diagram for the purple play category. The purple play category has five down lineman (squares) and one safety (a FS) in the backfield.

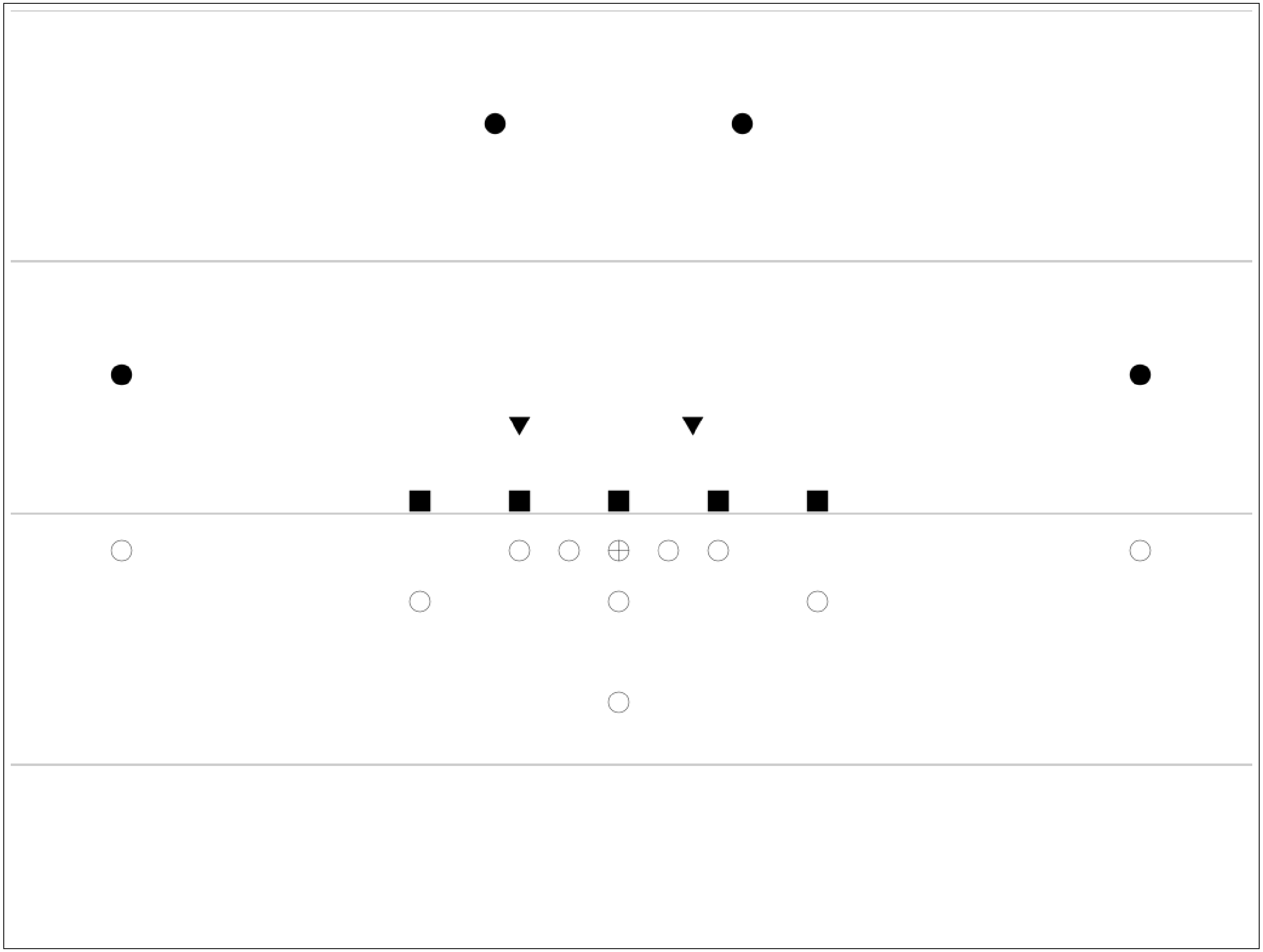

Figure 27: Prototype play diagram for the yellow play category. The yellow play category has five down lineman (squares) and two safeties in the backfield.

The cyan play is categorized by the presence of an additional player, a nickel back (NB), who can line up either near the cornerbacks or near the safeties. Additionally, the *cyan* play category takes a variable number of down linemen (3 or 4). Thus, to categorize example instances of the inconsistent category, participants have to learn a new feature to focus on: a new player position (NB), which can be either near a CB or in the backfield (near the other safeties). Participants also have to learn to ignore a feature that was previously attended: the number of down linemen. Figures 28 and 29 show prototypes for the cyan play category, illustrating that the number of DLs and the particular location of the NB is irrelevant.

This novel categories task was designed to assess positive and negative transfer. The consistent and inconsistent plays can be used to determine if one practice schedule facilitates developing a superordinate categorization strategy. Suppose participants in one schedule perform worse on the inconsistent play, relative to the consistent plays. This would suggest that they have learned to attend to those features critical for acquisition plays (number of squares and position of circles) but have difficulty ignoring those features and attending to a new critical feature: number of circles. If, on the other hand, participants in one schedule perform equally well on the consistent and inconsistent plays, then it suggests that they have learned to identify critical features in novel categories, independent of whether those critical features are the same or different from features critical for acquisition training.

After finishing the novel categories task participants completed another paper and pencil recall test. This recall test asked them to describe and draw the three play categories from the transfer task (i.e., cyan, purple, and yellow.)

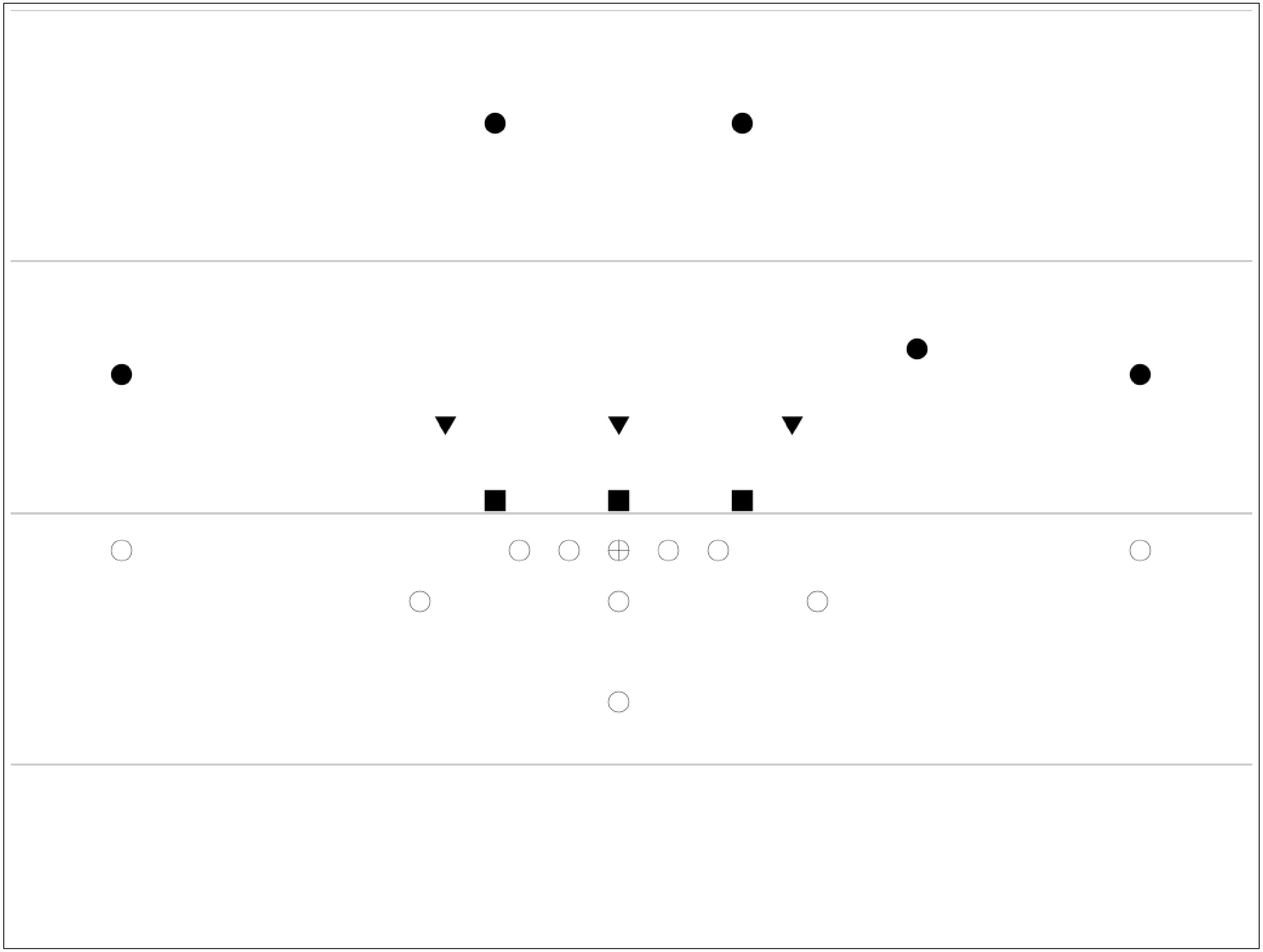

Figure 28: Prototype (#1) play diagram for the cyan play category. The cyan play category has five defensive backs: two CBs, two Ss, and one NB. The cyan category is the only play category with a NB.

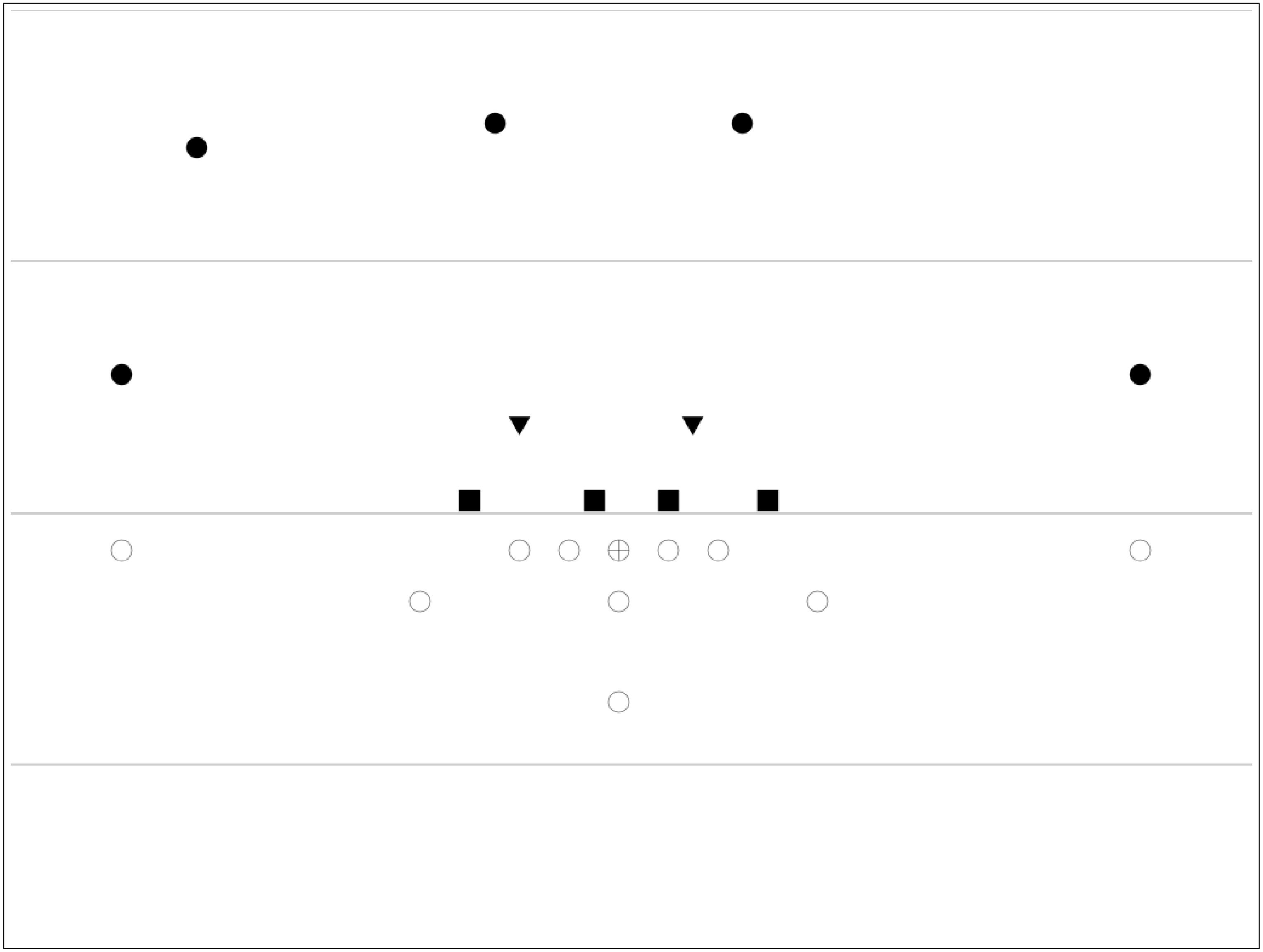

Figure 29: Prototype (#2) play diagram for the cyan play category. The cyan play category has five defensive backs: two CBs, two Ss, and one NB. The cyan category is the only play category with a NB.

4.3 Results

4.3.1 Structuring acquisition blocks

The data were separated into positive example instance trials and negative example instance trials. The positive example instance trials were restructured using the same procedure described in Experiment 1 to create blocks of 16 trials, with four trials for each play category.

The negative example instance trials were not restructured. Negative example instance trials were created such that each acquisition block has three negative examples, each from a different play category. Therefore, the reason for restructuring the positive example instances (i.e., to structure acquisition blocks so that they sample all four categories) is unnecessary for negative example instances: these blocks already contain a mix of play categories.

Errors were analyzed together for all trials (i.e., both positive and negative example instance trials). The errors on all trials were not restructured.

4.3.2 Dependent Measures

4.3.2.1 Accuracy

Each response was scored for categorization accuracy (i.e., correct responses and errors). Correct responses consist of hits and correct rejections, and errors consist of misses and false alarms. Incorrect trials (errors) were repeated and only correct trials were used to analyze the response time data.

4.3.2.2 Response Time

Response time of correct categorizations (i.e., response time; RT) was analyzed separately for positive and negative example instances. They were analyzed separately because they are conceptually different: the negative example instances were introduced primarily as catch trials and had only three trials per block.

Each participant's median response time for a block (i.e., 16 trials for positive example instances or 3 trials for negative example instances) was calculated. Median response times were used to reduce the influence that a single, long RT might have on the measure of central tendency for the block.

4.3.3 Session 1

4.3.3.1 Response Time on Positive Examples

Response time on each block was calculated for positive example trials; Figure 30 shows RT for the three practice schedule conditions. Early during acquisition, median response time was longer for the random schedule than either the blocked or blocked-repeated schedules. By the middle of acquisition, however, RT for the three practice schedules seemed to converge.

Performance on the initial and final blocks RT on the initial and final blocks was analyzed using two univariate ANOVAs with practice schedule as a between subjects factor. RT on the initial block was log transformed to remove positive skew in the distribution. Practice schedule had a significant effect on RT during the initial block, $F_{(2,36)} = 12.42$, $p < .001$, $MSE = 0.16$. Both the blocked schedule and the blocked-repeated schedule conditions had lower RT than the random schedule condition (mean log difference = 0.73, $p < .001$ and mean log difference = 0.68, $p < .001$, respectively). There was no significant difference between the blocked and blocked-repeated schedule conditions (mean log difference = 0.05, $p = .93$). On the final block, practice schedule did not have a significant effect on RT, $F_{(2,36)} = 1.60$, $p = .22$, $MSE = 59843.61$. Learning curves were fit to the data to explore the effect of practice schedule over the entire acquisition session.

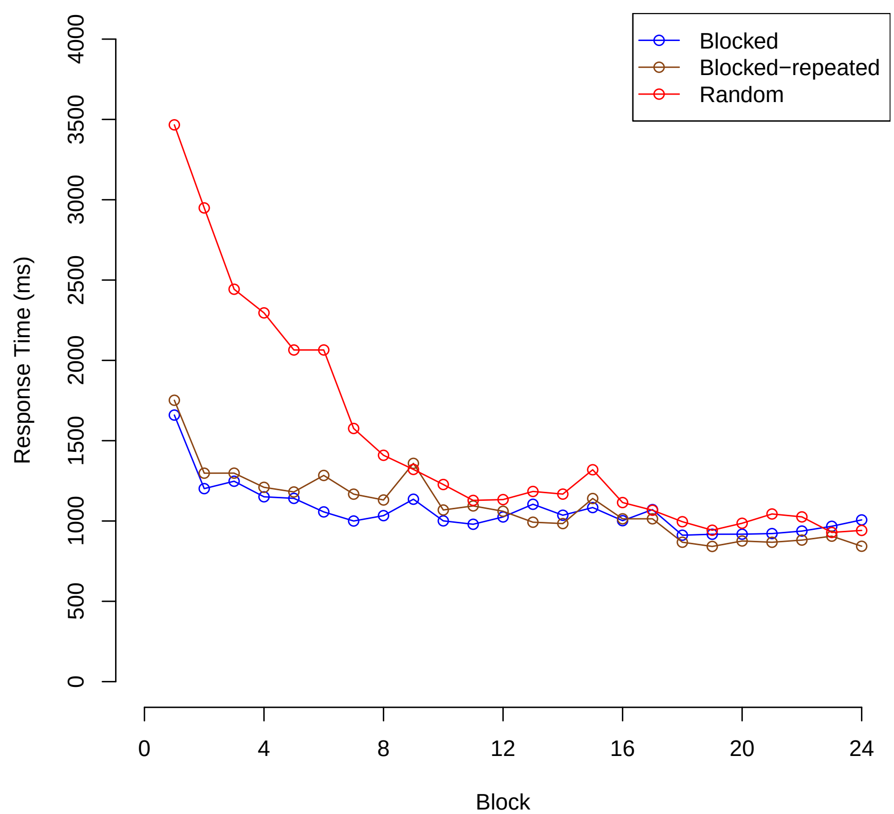

Figure 30: Response time for correct responses to trials with positive example instances during acquisition, as a function of practice schedule.

4.3.3.2 Fitting Learning Curves

Response time for correct trials with positive example instances was fit to a simple power curve using the same procedure as Experiment 1. Model fit and learning curve parameters are displayed in Table 30. Figure 31 displays these learning curves along with the empirical data to which they are fit. The learning rate parameter (α) and the initial performance parameter (β) are highest for the random schedule condition and lowest for the blocked schedule condition. The random schedule condition has a larger β than either the blocked or blocked-repeated conditions, approximately 10 s longer than both the blocked and blocked-repeated conditions. In addition, the random schedule has a larger α, more than double the blocked and blocked-repeated conditions.

As seen in Figure 31, the combined effect of the higher initial performance and higher learning rate results in the random schedule condition's RT approaching and, by the end of training, reaching the other schedule conditions' RT. Additionally, if performance is extrapolated to further training blocks the random schedule condition's RT continues to drop. This suggests that with additional acquisition trials the random schedule condition's RT could continue to decrease below the blocked and blocked-repeated schedule condition RTs, assuming performance does not reach an asymptote. Performance in the blocked-repeated schedule condition closely matches performance in the blocked schedule condition, though it does have a larger β and α than the blocked schedule condition.

4.3.3.3 Response Time on Negative Trials

The pattern of response times for negative example trials approximates the pattern of response times for positive example trials, although during early acquisition blocks there is less separation between the blocked and random schedule conditions (relative to the separation between these two conditions on the positive example trials). Early

Table 30: Experiment 3. Model fit and parameters derived from fitting response time on positive example instance trials for each practice schedule to a simple power law.

Acquisition schedule	R^2	Parameters β	α
Blocked	.76	2099.95	.14
Blocked-repeated	.82	2938.75	.20
Random	.95	13702.75	.45

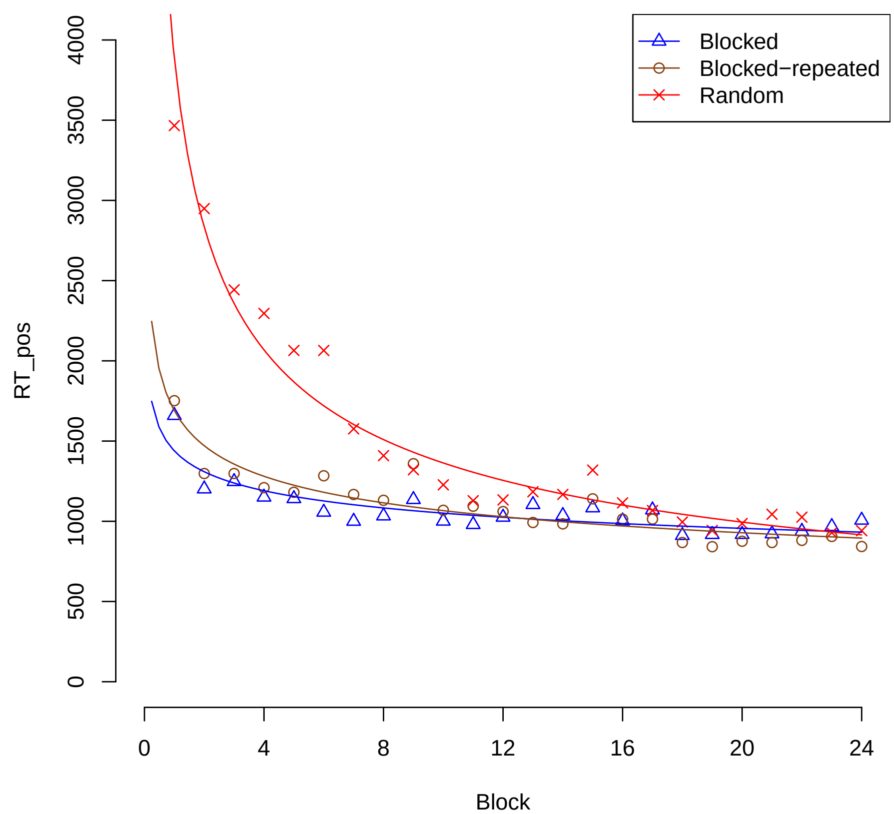

Figure 31: Response time for correct responses to trials with positive example instances during acquisition, with curves based on fitting data to a simple power law, as a function of practice schedule.

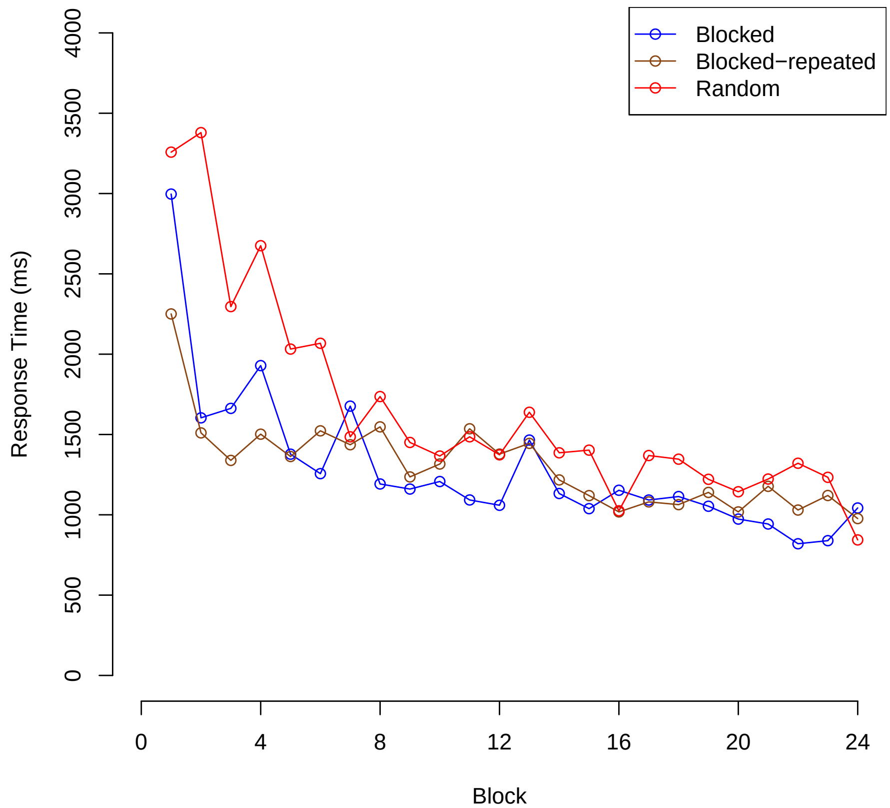

Figure 32: Response time for correct responses to trials with negative example instances during acquisition, as a function of practice schedule.

during acquisition, response times on negative example trials were higher for the random schedule than either the blocked or blocked-repeated schedules (see Figure 32). By the middle of acquisition, however, response time for the three acquisition schedule conditions appear to converge. These data suggest that, with training, participants in all acquisition schedule conditions become faster at responding to negative example trials.

4.3.3.4 Error

Median error per block was calculated for each practice schedule (see Figure 33). Median error for all acquisition schedule conditions was similar. For the majority of

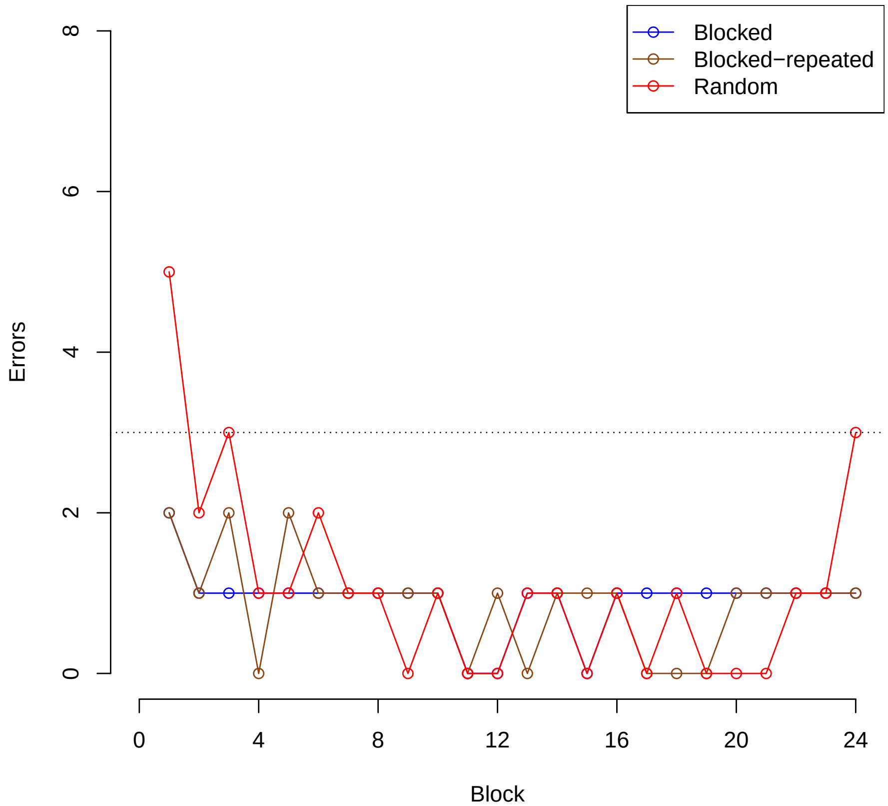

Figure 33: Errors on trials with positive and negative example instances during acquisition, as a function of practice schedule.

acquisition blocks, all conditions had median error less than three errors per block (approximately 85% accuracy or better). Because there were three negative example instances per block, if participants were not able to identify the negative example instances then one would expect error rates equal to or greater than three. These error data suggest that participants were able to learn the rules by which to assign diagrams to play categories. Additionally, errors do not increase with additional acquisition training, indicating that the decrease in response times that occurs with training is not due to participants trading speed for accuracy.

4.3.3.5 Category rules

At the conclusion of acquisition training participants completed a post-task questionnaire in which they listed the rules that determined category membership and drew a prototype play diagram (one participant did not complete the questionnaire). The verbal rules and drawn plays were assessed together to code whether the participant had "learned" the rules that govern category membership for each of the four plays. Categorization score could range from zero to four. Scores were negatively skewed, with many participants scoring a four. Participants in the blocked schedule condition had the best performance (median = 4, 25% quantile = 4, 75% quantile = 4), followed by participants in the random condition (median = 4, 25% quantile = 3, 75% quantile = 4), and participants in the blocked-repeated condition (median = 4, 25% quantile = 2.5, 75% quantile = 4). However, the categorization scores did not significantly differ among conditions, $\chi^2(8, N = 38) = 8.73$, $p = .37$.

4.3.3.6 Prior knowledge

The football knowledge pretest (see Appendix D) was scored to assess prior knowledge. The questions were scored for accuracy and were aggregated into two measures: general (Questions 1, 2, and 4) and specific (Questions 3 and 5). These measures were highly correlated ($r = .81$, $p < .001$). An aggregate *prior knowledge* measure was created as follows: general and specific prior knowledge were individually transformed into proportion of correct knowledge demonstrated (0–1.0) and then combined into a weighted average where specific knowledge was weighted twice as much as general knowledge. This aggregate measure was used to preserve differencess at the low and high ends of the scale. Scores on the prior knowledge measure were low (median = .13; 25% quantile = .06, 75% quantile = .28). The distribution of prior knowledge was non-normal (positively skewed), with many participants scoring 0 (n = 8, 21%). An arc sine square-root tranformation (appropriate for proportions and data with 0's) was used to transform the data closer to a normal distribution. The relation between prior knowledge and other dependent variables in Session 1 is presented in Table 31.

Given the high (though not significant) correlation between prior knowledge and intial response time, an additional analysis was conducted to assess the effect of practice schedule on response time during the initial block, with prior knowledge included as a covariate. As earlier, response time was log transformed prior to analysis. Prior knowledge did not have a significant effect on RT during the initial block, $F_{(1,33)} = 1.71$, $p = .20$, $MSE = 0.16$, nor did it have a significant interaction with acquisition schedule, $F_{(2,33)} = 0.18$, $p = .84$. With prior knowledge included in the model, acquisition schedule still significantly affected RT during the initial block, $F_{(2,33)} = 9.28$, $p < .001$. Including prior knowledge in the model improved the fit ($R^2_{schedule} = .41$ vs. $R^2_{scheduleXprior} = .44$) but resulted in a loss of degrees of freedom, reducing overall fit and model parsimony ($R^2_{schedule} = .38$ vs. $R^2_{scheduleXprior} = .36$).

4.3.4 Session 2

Six participants did not return for Session 2 and were excluded from further analysis. Cell sizes for the retention data are displayed in Table 32.

4.3.4.1 Skill retention

Data analysis procedure The data analysis procedure followed the approach used for the acquisition data: median response times on positive example trials were calculated for each participant.

Participants completed two retention blocks: a blocked retention block and a random retention block. The order of the retention blocks was counterbalanced across participants. This counterbalance variable, retention order, was included in the statistical models in order to account for any differences that might arise from completing the blocked retention block first or second.

Table 31: Experiment 3. Correlations between prior knowledge, performance measures, and the digit substitution task.

Variable	Prior	Initial RT	Final RT	Initial error	Final error	Rules	Digit
Prior	1.00	$-0.31^{§}$	-0.17	-0.03	-0.15	-0.01	-0.33^{*}
Initial RT		1.00	0.43^{**}	0.54^{***}	0.38^{*}	0.09	-0.16
Final RT			1.00	0.35^{*}	0.02	0.03	-0.07
Initial error				1.00	$0.29^{§}$	-0.13	-0.24
Final error					1.00	-0.27^{*}	0.06
Rules						1.00	0.08
Digit							1.00

$^{§}p < .10$ $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Table 32: Experiment 3. Number of participants (during Session 2, retention) in each acquisition schedule and retention order cell.

	Retention order		
Acquisition schedule	Blocked first	Random first	Total
Blocked	6	6	12
Blocked-repeated	6	6	12
Random	4	5	9
Total	16	17	33

Response time on Positive Trials Response time was calculated separately for the blocked and random retention blocks. Response time descriptive statistics are shown in Table 33. RT was analyzed separately for each retention block, using two-way ANOVAs with acquisition schedule and retention order as between subject factors.

In the blocked retention block, there was no significant effect of acquisition schedule, $F_{(2,27)} = 0.89$, $p = .42$, $MSE = 210994.40$, retention order, $F_{(1,27)} = 3.86$, $p = .06$, nor was there a significant interaction between acquisition schedule and retention order, $F_{(2,27)} = 0.68$, $p = .51$. In the random retention block, there was no significant effect of acquisition schedule, $F_{(2,27)} = 1.60$, $p = .22$, $MSE = 525076.90$, retention order, $F_{(1,27)} = 0.53$, $p = .47$, nor was there a significant interaction, $F_{(2,27)} = 0.04$, $p = .96$.

Errors Errors were calculated separately for the blocked and random retention blocks. In both the blocked and random retention blocks the distribution of errors was positively skewed, with the majority of participants making less than three errors per block. Because of the non-normal distribution of errors, error bins were created which corresponded to different conceptual amounts of error: 0 errors (100% correct categorizations), 1–3 errors (85–95% correct categorizations; the range one would expect if a participant was correct on all positive examples and incorrect on only

Table 33: Experiment 3. Mean RT (and standard deviation) for positive example instance trials during retention, as a function of acquisition schedule, retention order, and retention block schedule.

	Retention order		
Acquisition schedule	Blocked first	Random first	Average
	Blocked retention block		
Blocked	1796.08 (690.10)	1246.58 (247.15)	1521.33 (571.48)
Blocked-repeated	1477.17 (613.72)	1234.33 (350.91)	1355.75 (493.21)
Random	1293.50 (132.67)	1198.30 (338.55)	1240.61 (257.73)
	Random retention block		
Blocked	1853.50 (637.74)	1943.08 (388.88)	1898.29 (505.76)
Blocked-repeated	1853.92 (817.38)	2096.00 (1101.94)	1974.96 (933.60)
Random	1315.50 (77.92)	1548.70 (699.52)	1445.06 (511.91)

negative examples), and greater than 3 errors (less than 85% correct categorizations; the range one would expect if a participant was incorrect on several positive and negative examples). Table 34 displays the frequency of participants in these error bins on the blocked and random retention blocks, for each practice schedule. A two-way Pearson's chi-squared test (with Yates's correction for continuity) was conducted to test whether there was a difference in the distribution of errors among acquisition schedule conditions. For the blocked retention blocks, there was no significant difference among acquisition schedule conditions, $\chi^2(4, N = 33) = 0.28$, $p = .99$. Likewise, for the random retention blocks, there was no significant difference among acquisition schedule conditions, $\chi^2(4, N = 33) = 2.15$, $p = .70$. These data suggest that the majority of the participants in all conditions made few incorrect errors when categorizing diagrams on retention trials.

4.3.4.2 Recall of Category Rules

Prior to completing the retention trials, participants completed the same questionnaire from Session 1 in which they listed the rules that determined category membership and drew a prototype play diagram. As in Session 1, scores were negatively

Table 34: Experiment 3. Frequency for error range bins for all trials during retention, as a function of acquisition schedule, and retention block schedule.

	Retention order		
Acquisition schedule	0	1–3	>3
Blocked retention block			
Blocked	5	6	1
Blocked-repeated	4	7	1
Random	3	5	1
Total	12	18	3
Random retention block			
Blocked	5	5	2
Blocked-repeated	3	6	3
Random	5	3	1
Total	13	14	6

skewed, with many participants scoring a four. Across practice schedule conditions scores were similar (blocked schedule condition: median = 4, 25% quantile = 2.75, 75% quantile = 4; blocked-repeated schedule condition: median = 4, 25% quantile = 2.75, 75% quantile = 4; random schedule condition: median = 4, 25% quantile = 3, 75% quantile = 4). These categorization scores were not significantly different between condition, $\chi^2(8, N = 33) = 0.92$, $p < 1.00$.

4.3.4.3 Category assignment task

The category assignment task had two blocks of 16 trials (four trials per category). Both blocks were ordered according to a random schedule. The double pattern transfer task was analyzed for differences among acquisition practice schedules. Median RT for correct trials was calculated for each block. Mean median RT (and standard deviations) for each acquisition schedule condition and transfer block are displayed in Table 35.

A repeated measures analysis was conducted with acquisition schedule as a between subject factor and block as a repeated measure. Block had a significant effect on response time (see Table 35), $V = 0.28$, $F_{(1,30)} = 11.86$, $p = .002$, with errors

Table 35: Experiment 3. Mean median response time (and standard deviation) for correct trials on the category assignment task, as a function of acquisition schedule and block.

	Block		
Acquisition Schedule	1	2	Average
Blocked	2749.96 (1319.96)	2182.58 (632.69)	2466.27 (935.71)
Blocked-repeated	2154.92 (671.09)	1858.04 (564.69)	2006.48 (578.49)
Random	2573.06 (552.95)	2338.56 (687.55)	2455.81 (594.99)
Average	2485.33 $(948.43)_a$	2107.11 $(637.07)_b$	2296.22 (744.31)

Note. Means in each row that share a subscript do not differ significantly ($p < .05$).

decreasing from Block 1 to Block 2. Acquisition schedule did not have a significant effect on response time, $V = 0.09$, $F_{(2,30)} = 1.47$, $p = .25$, nor was there a significant interaction between acquisition schedule and transfer block, $V = 0.05$, $F_{(2,30)} = 0.87$, $p = .43$.

4.3.4.4 Novel categories task

The novel categories transfer task had six blocks of 19 trials (16 positive and 3 negative example instances per block). Three of the blocks were ordered according to a blocked practice schedule and three blocks were ordered according to a random practice schedule. All participants completed both blocked and random retention blocks; the order of blocks (i.e., blocked first or random first) was counterbalanced across participants. Each participant's Median RT for correct, positive example instance trials was computed for each block. The blocked retention blocks were restructured so that each block had five trials of each of the three play categories; this required dropping the last trial of each block.

Figure 34 shows median RT for each block as a function of acquisition practice schedule. Both the blocked and random transfer blocks are displayed. Practice schedule did not affect median RT for either the blocked or the random retention blocks.

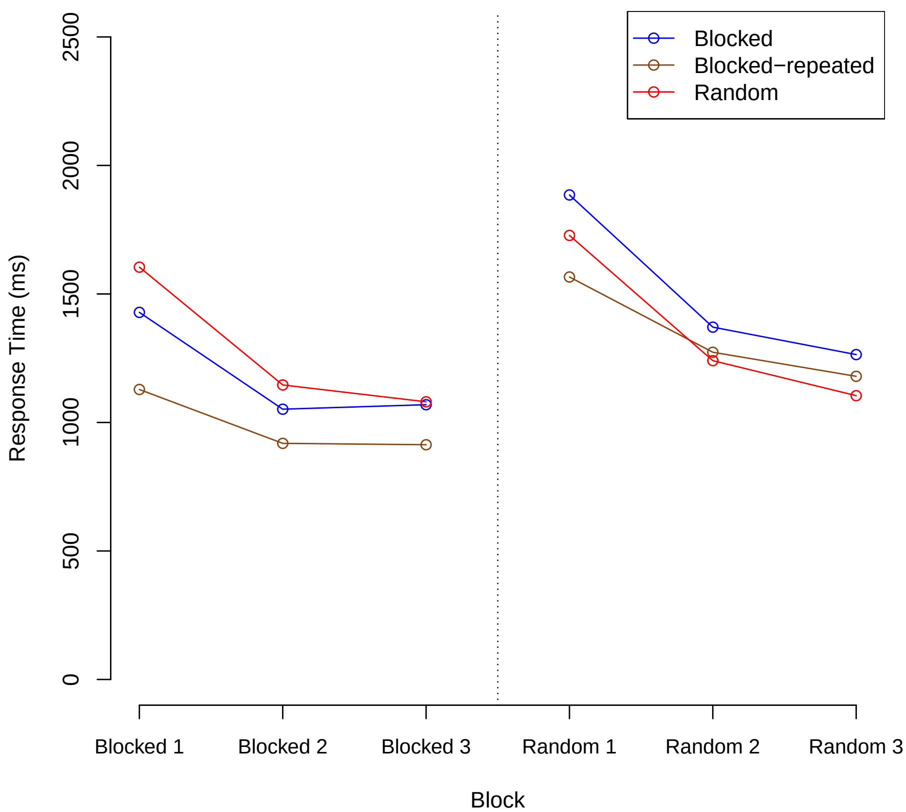

Figure 34: Response time for correct responses to trials with positive positive example instances during the novel categories transfer task, as a function of practice schedule.

4.4 Discussion

During the early acquisition training blocks, the random schedule condition has higher response times for both the positive and negative example trials and higher error rates on all trials. This difference disappears, however, by the end of acquisition training. Examining response time and errors during the course of acquisition training indicates that the random schedule condition's response times seem to converge with the blocked and blocked-repeated schedule conditions' response times by the middle of training (i.e., around block 10). Furthermore, fitting simple power curves to response time performance on positive example instance trials suggests that if acquisition training was continued, response times for the random schedule condition would continue to decrease, possibly surpassing response times for the blocked (and blocked-repeated) schedule conditions. In terms of response times and error, performance in the blocked-repeated schedule condition is similar to the blocked schedule condition.

In this experiment, skill retention does not seem to be strongly affected by the acquisition practice schedule. In both blocked and random retention blocks, the acquisition practice schedule does not significantly affect median RT or error rates during retention. There is a slight, non-significant trend for the random acquisition schedule condition to have faster response times than the blocked and blocked-repeated schedule conditions when retention trials are ordered according to a random schedule, but this trend is not reliable. Likewise, errors during retention are not affected by the acquisition practice schedule. Together, these results suggest that, relative to the blocked or blocked-repeated acquisition schedules, a random acquisition schedule does not result in faster categorization or more error-free categorization trials after a 48 hour delay. The category assignment and novel categories transfer tasks also show the same non-significant effects of practice schedule on RT.

As with the motor learning task experiments (Experiments 1 and 2), a different pattern of results is seen in acquisition and retention/transfer. In acquisition, the blocked and blocked-repeated schedule conditions have better performance than the random condition early in training. This difference decreases (and might eventually reverse) with extended training. On the other hand, there is no effect of acquisition schedule during retention or transfer.

Early in training, when participants are concentrating on learning rules, a blocked schedule reduces response time and errors, suggesting that the blocked schedule facilitates induction of category rules. As both groups learn these rules, and skilled performance relies more on speed of classification, the RT differences among schedules disappear. On retention and transfer, classification speed is again important and there are not reliable differences between acquisition schedule conditions.

CHAPTER V

EXPERIMENT 4

Experiment 4 was conducted to evaluate the effects of practice schedule and amount of practice on acquisition, retention, and transfer of the perceptual categorization task.

5.1 Method

5.1.1 Participants

Ninety-one participants over the age of 17 were recruited. Seventy-three participants were recruited from the Georgia Tech psychology participant pool and participated for course credit. Eighteen participants were Georgia Tech students that participated for pay ($12 per hour)[1]. All participants were recruited with the exclusion criterion that they had not "played on a football team in high school or college".

Fifteen participants were excluded prior to data analysis for three reasons: software crashed during training (N = 13), played football in high school (N = 1), and ignored instructions on how to perform the task (N = 1). Three additional participants were excluded because their data fit the exclusion criterion defined in Experiment 3 (see Appendix C). All three of these participants were in the random condition.

Excluding these participants resulted in 73 participants; cell sizes are displayed in Table 36. Participants' mean age was 20.2 years ($SD = 2.2$). Of the 76 participants, 31 were male, 31 were female, and 11 did not answer.

---

[1] Of these paid participants, three were in the low amount condition, six were in the medium amount condition, and nine were in the high amount condition.

Table 36: Experiment 4. Number of participants (during Session 1, acquisition) in each acquisition schedule, acquisition schedule version, and amount of practice cell.

	Version				
Acquisition schedule	1	2	3	4	Total
Low Amount					
Blocked	5	3	3	3	14
Random	6	5	–	–	11
Medium Amount					
Blocked	3	3	2	3	11
Random	7	6	–	–	13
High Amount					
Blocked	3	5	4	3	15
Random	3	6	–	–	9

Note. The random schedule condition had only two versions.

5.1.2 Materials

The same categorization task and stimuli from Experiment 3 were used. Additionally, the same football knowledge questionnaire was used to determine participants' familiarity with football.

5.1.3 Design

5.1.3.1 Practice schedule

Two practice schedules were used: blocked and random. In the blocked schedule, each training block used one play category and participants classified diagrams as positive or negative examples of the category. In the random schedule, each training block used all four play categories and participants classified diagrams as positive or negative examples of each category. Experiment 4 used the same schedule versions that were used in Experiment 3.

As in Experiment 3, each acquisition block consisted of 19 trials: 16 positive examples and 3 negative examples. Experiment 4 used the same method for distributing play categories and play diagrams as Experiment 3.

5.1.3.2 *Amount of practice*

Participants were assigned to one of three amount of practice conditions: low, medium, or high. The low amount had 4 blocks (76 trials: 64 positive example instances and 12 negative example instances); the medium amount had 8 blocks (152 trials: 128 positive example instances and 24 negative example instances); the high amount had 16 blocks (304 trials: 256 positive example instances and 48 negative example instances). These amounts were selected based on the learning curves fit from the Experiment 3 data, comparisons to existing research using the practice schedule manipulations, and the practical constraints of creating blocks of trials that could be equally sequenced with four task categories.

Specifically, the low amount was chosen to correspond to a location where there was still rapid improvement in learning and a large separation between performance in the blocked and random schedule conditions. The medium condition was placed where performance improvement began leveling off and the high condition was placed where the performance for the two schedule conditions began to converge. Additionally, blocks were chosen such that positive and negative examples could be distributed consistently across blocks in all low, medium, and high amount conditions, and where the high amount condition was twice the number of trials as the medium amount.

5.1.3.3 *Retention interval*

Participants completed two sessions: Session 1 (acquisition), and 48 hours later, Session 2 (retention and transfer).

5.1.4 Procedure

The same procedure as Experiment 3 was used. The only change to the procedure resulted from the amount of practice manipulation. In Experiment 3 participants completed paper and pencil tasks during the breaks between epochs, in order to minimize fatigue effects. In Experiment 3 an epoch was defined as 8 blocks. Experiment 2 epochs also had 8 blocks, which means that only the high amount condition had more than one epoch. Regardless of condition, however, all participants completed the same paper and pencil tasks. In the low amount and medium amount conditions participants completed these tasks after their final acquisition block (Blocks 4 and 8, respectively). In the high amount condition participants completed the first set of tasks after Epoch 1 (Block 8) and the second set after Epoch 2 (Block 16).

5.2 Results

5.2.1 Structuring acquisition blocks

Trials in the blocked schedule condition were restructured so that each block included four trials of each category. Participants' median response time on each block was calculated for positive example trials in that block, following the procedure described in Experiment 3.

5.2.2 Session 1

5.2.2.1 Response Time on Positive Examples

Median response time during acquisition is graphed for the low, medium, and high amount conditions in Figures 35, 36, and 37, respectively. Figure 38 shows response time on positive example trials for all practice schedule and amount of practice conditions.

Performance on the final block To assess the effect that acquisition schedule and amount of practice had at the conclusion of training, I examined performance during the final acquisition block. Means and standard deviations for each acquisition schedule by amount of practice condition are displayed in Table 37. RT on the final block was log transformed to remove positive skew in the distribution.

Performance on the final acquisition block was modeled using a full factorial model with acquisition schedule and amount of practice. There was a significant main

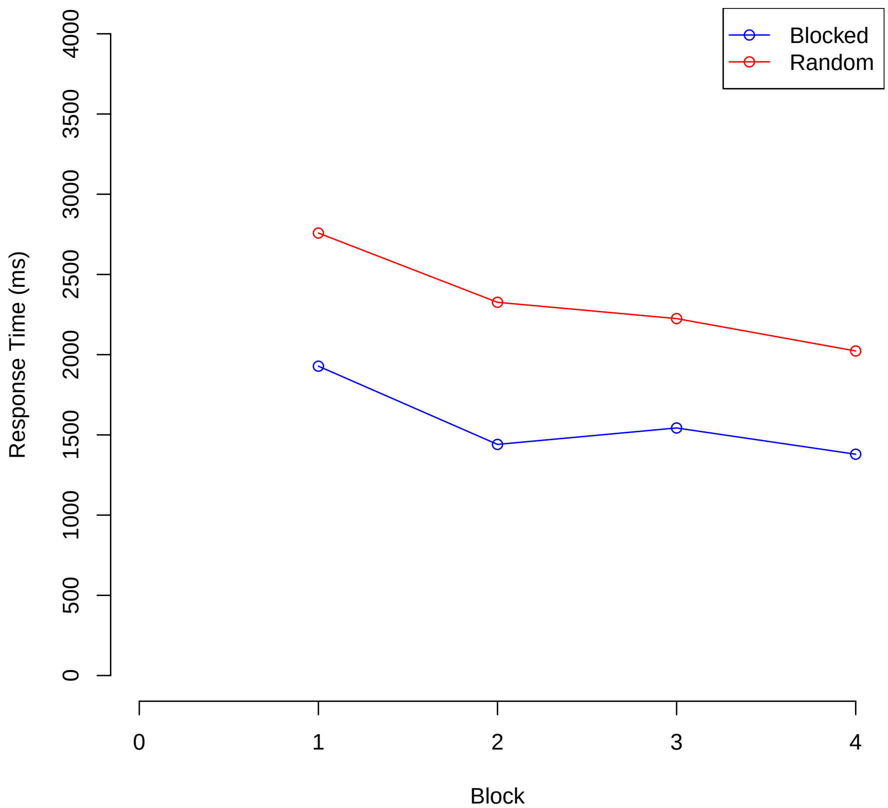

Figure 35: Response time for correct responses to trials with positive example instances during acquisition for the low amount of practice conditions, as a function of practice schedule and amount of practice.

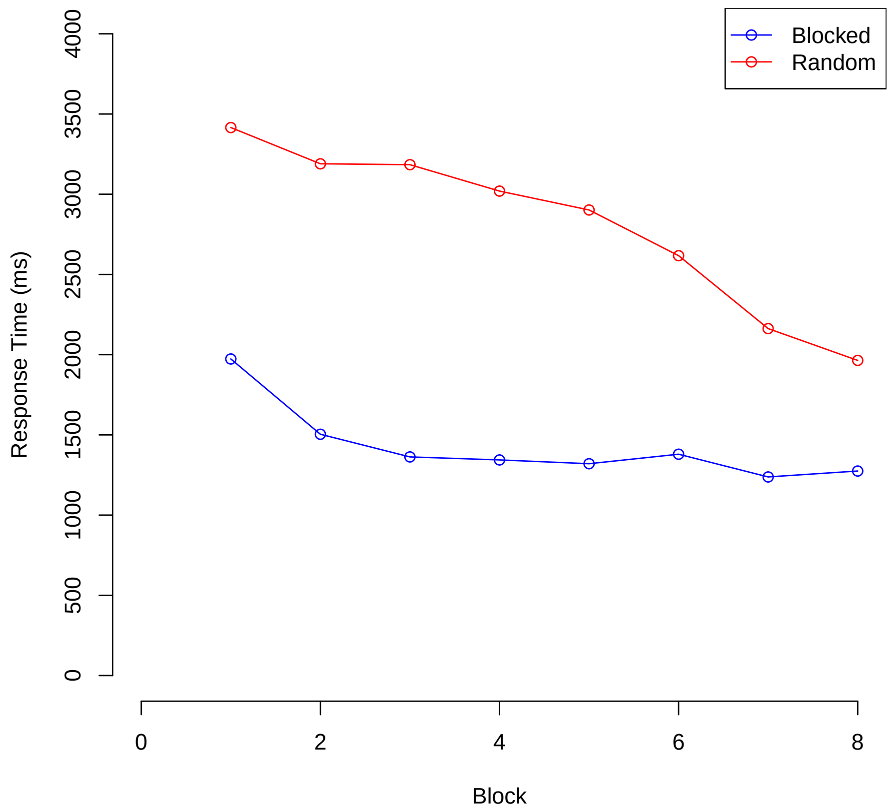

Figure 36: Response time for correct responses to trials with positive example instances during acquisition for the medium amount of practice conditions, as a function of practice schedule and amount of practice.

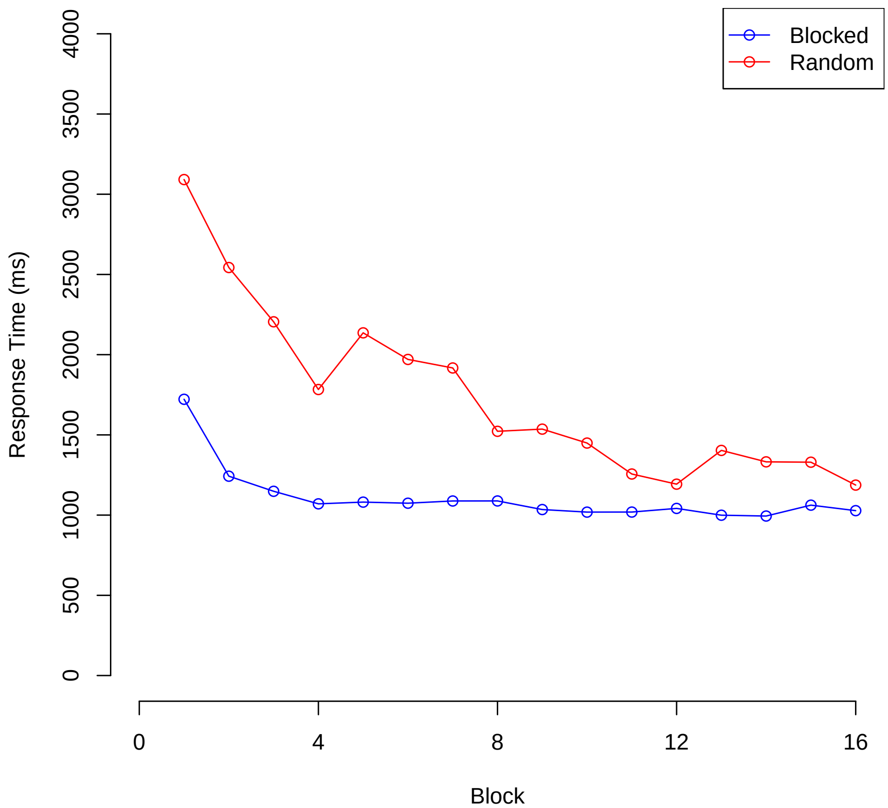

Figure 37: Response time for correct responses to trials with positive example instances during acquisition for the high amount of practice conditions, as a function of practice schedule and amount of practice.

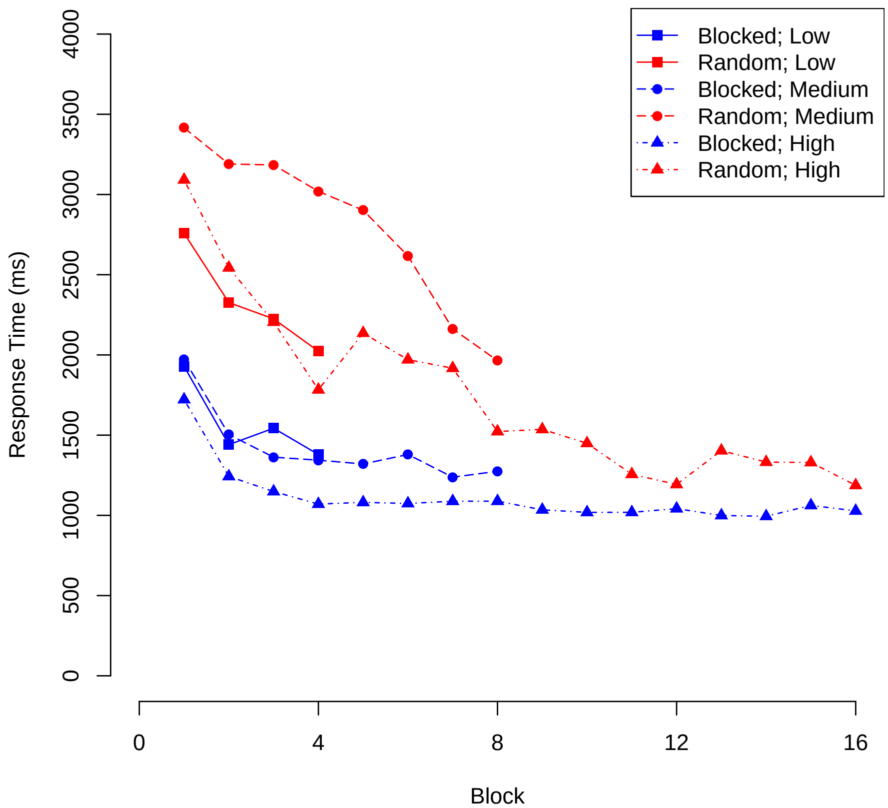

Figure 38: Response time for correct responses to trials with positive example instances during acquisition, as a function of practice schedule and amount of practice.

effect of acquisition schedule, $F_{(1,67)} = 14.17$, $p < .001$, $MSE = 0.14$, and there was a significant main effect of amount of practice, $F_{(2,67)} = 6.45$, $p = .003$. The interaction between acquisition schedule and amount of practice was not significant, $F_{(2,67)} = 0.91$, $p = .41$. The high amount of practice conditions had lower RT than both the low and the medium amount of practice conditions (mean log difference = 0.39, $p = .004$ and mean log difference = 0.36, $p = .01$, respectively). There was no significant difference between RTs for the low amount of practice and the medium amount of practice conditions (mean log difference = 0.03, $p = .96$).

5.2.2.2 Response Time on Negative Trials

The pattern of response time for negative example trials across practice schedules was more noisy than response time on negative trials (see Figure 39). This might arise, in part, because the negative example blocks have only 3 trials per block whereas the positive example blocks have 16 trials per block. On negative example trials, the blocked and random schedule conditions had similar response times except for the random, medium amount condition. This condition also had higher response times in the positive examples condition. Overall, the negative example trial response time data shows that participants in all conditions improved their speed in classifying negative example instances during training.

5.2.2.3 Errors

Median error per block was calculated for each practice schedule and amount of practice condition (see Figure 40). During the first three blocks the random schedule conditions had a consistent pattern of higher median error than the blocked schedule conditions. After these initial blocks, however, all conditions had medians lower than three errors per block (approximately 85% accuracy or better). These error data suggest that participants in both acquisition schedule conditions were able to learn the rules by which to assign diagrams to play categories. Additionally, errors decreased

Table 37: Experiment 4. Mean median response time (and standard deviation) on positive example instance trials during the final acquisition block, as a function of acquisition schedule and amount of practice.

	Amount			
Acquisition schedule	Low	Medium	High	Average
Blocked	1379.46 (536.08)	1274.73 (575.49)	1027.83 (355.93)	1218.80 (500.36)
Random	2023.00 (613.20)	1964.19 (771.02)	1187.17 (292.98)	1771.88 (703.39)
Average	1662.62 (647.02)	1648.19 (759.80)	1087.58 (336.42)	1468.82 (657.52)

Note. Means in each column that share a subscript do not differ significantly ($p < .05$).

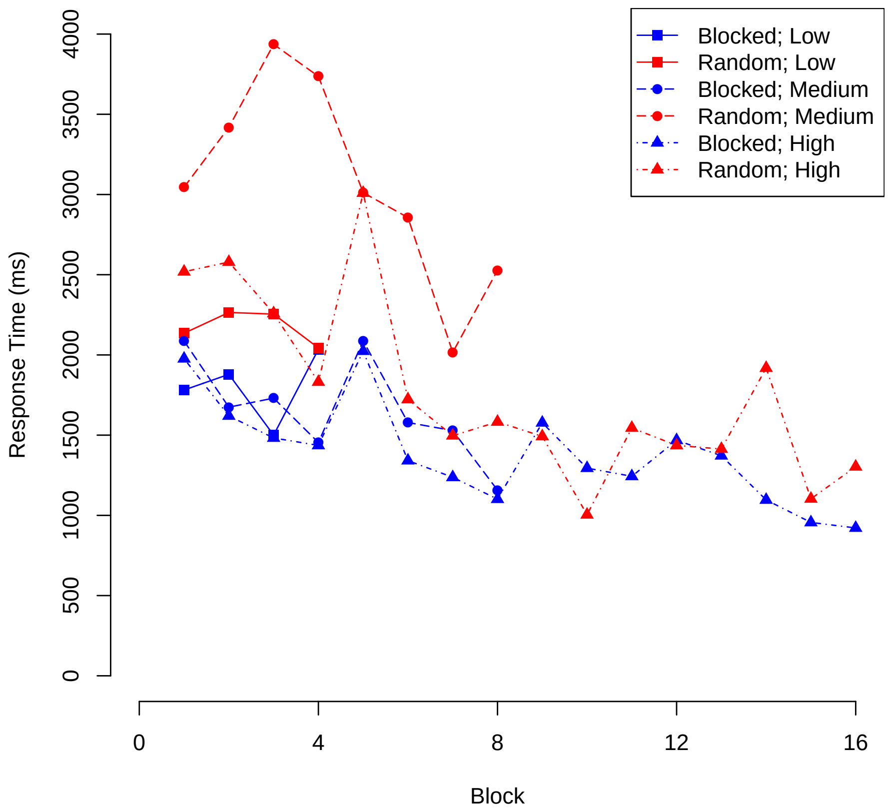

Figure 39: Response time for correct responses to trials with negative example instances during acquisition, as a function of practice schedule and amount of practice.

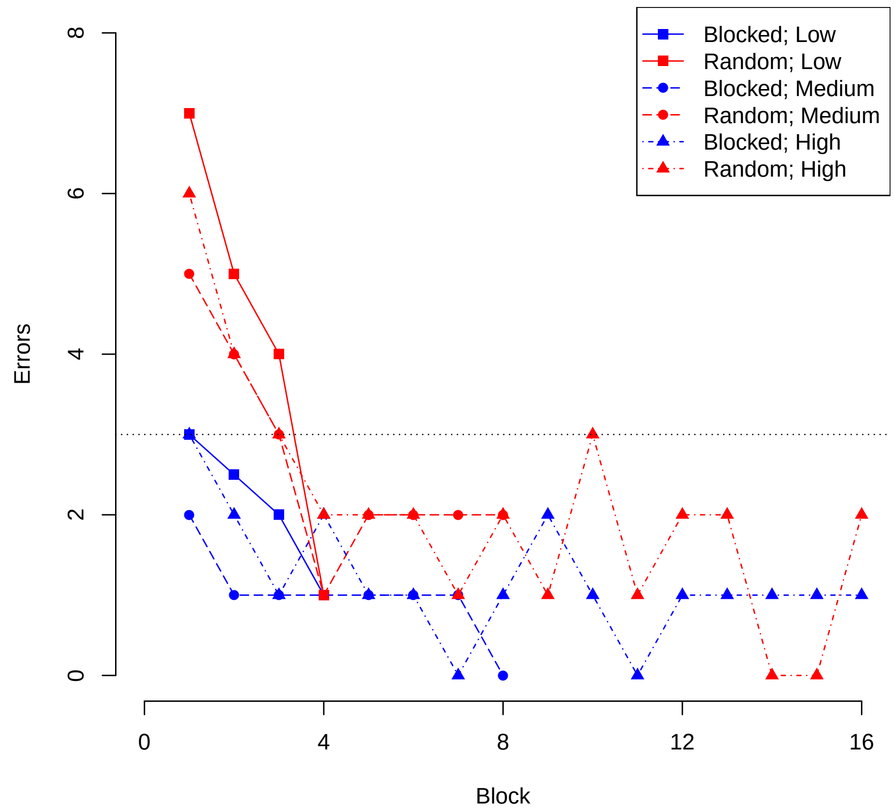

Figure 40: Errors on trials with positive and negative example instances during acquisition, as a function of practice schedule and amount of practice.

with additional acquisition training, indicating that the decrease in response times that occurs with training is not due to participants trading speed for accuracy.

5.2.2.4 Category Rules

At the end of Session 1 participants completed the post-task questionnaire in which they listed the rules that determined category membership and drew a prototype play diagram. Categorization score could range from zero to four. Scores were negatively skewed, with many participants scoring a four. Transforming the scores did not normalize the distribution. In order to explore the effects of each manipulation (acquisition schedule and amount of practice) each effect was analyzed separately.

Table 38: Experiment 4. Category rule scores in Session 1 as a function of acquisition schedule and amount of practice.

	Score				
Acquisition schedule	0	1	2	3	4
Low Amount					
Blocked	0	5	3	3	3
Random	1	2	4	1	3
Total	1	7	7	4	6
Medium Amount					
Blocked	0	0	5	0	6
Random	1	1	2	3	6
Total	1	1	7	3	12
High Amount					
Blocked	0	0	2	1	11
Random	0	0	3	1	5
Total	0	0	5	2	16

The acquisition schedule manipulation did not appear to affect errors, however, the amount of practice manipulation did (see Table 38). A chi-square test was used to further test the significance of this trend. Amount of practice affected rule learning scores, $\chi^2(8, N = 73) = 17.08$, $p = .03$, with the low amount of practice having the lowest scores.

5.2.2.5 Prior knowledge

As in Experiment 3, the football knowledge pretest was scored to assess prior knowledge. These general and specific knowledge measures were highly correlated ($r = .77$, $p < .001$), and an aggregate prior knowledge measure was created following the procedure described in Experiment 3. Scores on the prior knowledge measure were low (median = .13; 25% quantile = 0, 75% quantile = .43). The distribution of prior knowledge was non-normal (positively skewed), with many participants scoring 0 (n = 24; 33%). An arc sine square-root tranformation was used to transform the data closer to a normal distribution. The relation between prior knowledge and other dependent variables in Session 1 is presented in Table 39.

Table 39: Experiment 4. Correlations between prior knowledge, performance measures, and the digit substitution task.

Variable	Prior	Initial RT	Final RT	Initial error	Final error	Rules	Digit
Prior	1.00	-0.32^{**}	-0.11	$-0.21^{\S}$	-0.17	0.07	0.12
Initial RT		1.00	0.59^{***}	$0.20^{\S}$	-0.05	0.01	-0.05
Final RT			1.00	0.17	0.05	-0.15	-0.18
Initial error				1.00	0.07	-0.19	-0.03
Final error					1.00	-0.58^{***}	0.18
Rules						1.00	-0.05
Digit							1.00

$^{\S}p < .10$ $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Given the significant correlation between prior knowledge and intial response time, an additional analysis was conducted to assess the effect of practice schedule on response time during the initial block, with prior knowledge included as a covariate. The model tested the effects of practice schedule, amount of practice, and prior knowledge, as well as the second-order interactions between practice schedule and amount of practice, practice schedule and prior knoweldge, and amount of practice and prior knowledge. As in Experiment 3, response time was log transformed prior to analysis.

Prior knowledge did not have a significant effect on RT during the initial block, $F_{(1,63)} = 2.49$, $p = .12$, $MSE = 0.15$, nor did it have a significant interaction with practice schedule, $F_{(1,63)} = 0.14$, $p = .71$, nor with amount of practice $F_{(2,63)} = 0.83$, $p = .44$. Practice schedule significantly affected RT during the initial block, $F_{(1,63)} = 25.41$, $p < .001$. There was no significant effect of amount of practice, $F_{(2,63)} = 1.12$, $p = .33$ and there was no signficant interaction between practice schedule and amount of practice, $F_{(2,63)} = 1.18$, $p = .31$.

5.2.3 Session 2

Six participants did not return for Session 2 and and were excluded from further analysis. One additional participant was excluded from the retention analysis because of a software error in saving the data file. Cell sizes for the retention data are displayed in see Table 40.

5.2.3.1 Skill retention

Data Analysis Procedure The data analysis procedure followed the acquisition procedure, median response times for positive example instances were calculated for each participant.

Participants completed two retention blocks: a blocked retention block and a random retention block. The order of the retention blocks was counterbalanced across participants and this counterbalance variable was included in the statistical models.

Table 40: Experiment 4. Number of participants (during Session 2, retention) in each acquisition schedule, amount of practice, and retention order cell.

	Retention order		
Acquisition schedule	Blocked first	Random first	Total
	Low Amount		
Blocked	7	6	13
Random	5	6	11
Total	12	12	24
	Medium Amount		
Blocked	4	6	10
Random	6	6	12
Total	10	12	22
	High Amount		
Blocked	6	6	12
Random	5	4	9
Total	11	10	21

Response time on Positive Trials Response time was calculated separately for the blocked and random retention blocks. Response time descriptive statistics for the blocked retention blocks and the random retention blocks are shown in Tables 41 and 42, respectively.

RT was analyzed separately for each retention block. Each analysis used an ANOVA with acquisition schedule, amount of practice, and retention order as between subject factors. The models tested the main effects of each factor and the two-way interactions between acquisition schedule and amount of practice, and between acquisition schedule and retention order. In the blocked retention block there was not a significant effect of acquisition schedule, $F_{(1,59)} < 0.01$, $p = .98$, $MSE = 528529.00$, nor was there a significant effect of amount of practice, $F_{(2,59)} = 2.41$, $p = .10$ or retention order, $F_{(1,59)} = 3.62$, $p = .06$. Additionally, there was no significant interaction between acquisition schedule and amount of practice, $F_{(2,59)} = 0.78$, $p = .46$, nor was there a significant interaction between acquisition schedule and retention order, $F_{(1,59)} = 0.68$, $p = .41$.

Table 41: Experiment 4. Mean median response time (and standard deviation) on positive example instance trials during the blocked retention block, as a function of acquisition schedule, retention order, and amount of practice.

	Retention order		
Acquisition schedule	Blocked first	Random first	Average
	Low Amount		
Blocked	2089.50 (1071.95)	2078.67 (853.04)	2084.50 (936.89)
Random	2149.00 (561.66)	1437.00 (753.48)	1760.64 (740.48)
Average	2114.29 (861.64)	1757.83 (837.33)	1936.06 (850.61)
	Medium Amount		
Blocked	1912.00 (935.90)	1569.33 (538.09)	1706.40 (695.80)
Random	2080.00 (988.14)	1818.67 (403.99)	1949.33 (732.55)
Average	2012.80 (917.58)	1694.00 (471.96)	1838.91 (709.87)
	High Amount		
Blocked	1586.25 (548.88)	1301.33 (502.72)	1443.79 (523.41)
Random	1804.80 (901.59)	1268.88 (301.58)	1566.61 (721.33)
Average	1685.59 (699.15)	1288.35 (413.52)	1496.43 (602.23)
	Average		
Blocked	1870.12 (860.81)	1649.78 (695.39)	1756.80 (776.51)
Random	2015.56 (805.46)	1538.09 (562.80)	1776.83 (725.27)
Average	1940.64 (824.66)	1597.22 (629.51)	1766.37 (746.85)

Like the blocked retention block, the random retention block was not affected by any of the factors. There was not a significant effect of acquisition schedule $F_{(1,59)} = 0.21$, $p = .65$, $MSE = 787045.30$, amount of practice, $F_{(2,59)} = 0.71$, $p = .49$, nor retention order, $F_{(1,59)} = 1.20$, $p = .28$. Additionally, there was no significant interaction between acquisition schedule and amount of practice, $F_{(2,59)} = 0.07$, $p = .93$, nor was there a significant interaction between acquisition schedule and retention order, $F_{(1,59)} = 0.24$, $p = .63$.

5.2.3.2 Recall of Category Rules

At the beginning of Session 2 participants completed the post-task category rules questionnaire again to assess their recall of category rules. As with the Session 1 data, categorization scores were negatively skewed. Although the same trend of lower

Table 42: Experiment 4. Mean median response time (and standard deviation) on positive example instance trials during the random retention block, as a function of acquisition schedule, retention order, and amount of practice.

	Retention order		
Acquisition schedule	Blocked first	Random first	Average
Low Amount			
Blocked	2011.93 (1265.48)	2166.92 (1042.18)	2083.46 (1122.38)
Random	2212.50 (779.22)	1999.08 (769.77)	2096.09 (742.68)
Average	2095.50 (1051.17)	2083.00 (877.91)	2089.25 (947.16)
Medium Amount			
Blocked	2004.12 (1080.46)	2502.00 (1184.87)	2302.85 (1111.39)
Random	1916.00 (914.54)	2206.17 (780.36)	2061.08 (824.58)
Average	1951.25 (925.13)	2354.08 (968.92)	2170.98 (949.05)
High Amount			
Blocked	1688.83 (473.96)	2103.25 (757.35)	1896.04 (640.05)
Random	1652.80 (481.81)	1977.75 (942.83)	1797.22 (691.92)
Average	1672.45 (453.36)	2053.05 (786.87)	1853.69 (647.56)
Average			
Blocked	1896.06 (956.31)	2257.39 (966.16)	2081.89 (964.71)
Random	1926.41 (744.91)	2071.41 (768.10)	1998.91 (747.93)
Average	1910.77 (847.11)	2169.87 (870.59)	2042.25 (862.51)

Table 43: Experiment 4. Category rule scores in Session 2 as a function of acquisition schedule and amount of practice.

	Score				
Acquisition schedule	0	1	2	3	4
Low Amount					
Blocked	2	2	6	1	2
Random	0	2	7	0	2
Total	2	4	13	1	4
Medium Amount					
Blocked	0	1	3	0	6
Random	1	1	3	3	4
Total	1	2	6	3	10
High Amount					
Blocked	0	2	2	2	6
Random	0	1	1	1	6
Total	0	3	3	3	12

scores for the low amount condition was present in Session 1 (see Table 43), the effect of amount of practice was not significant, $\chi^2(8, N = 67) = 14.50$, $p = .07$.

5.2.3.3 Category assignment task

As in Experiment 3, the category assignment task had two blocks of 16 trials (four trials per category). Both blocks were ordered according to a random schedule. The double pattern transfer task was analyzed for differences among acquisition practice schedule and amount of practice conditions. Median RT for correct trials was calculated for each block. Mean median response times (and standard deviations) for each acquisition schedule condition and transfer block are displayed in Table 44).

A three-way repeated measures analysis was conducted with acquisition schedule and amount of practice as between subject factors and block as a repeated measure. The only significant effect was the effect of block, $V = 0.20$, $F_{(1,60)} = 14.73$, $p < .001$, with response time decreasing from Block 1 to Block 2 (see Table 44). There was no significant effect of acquisition schedule $V = 0.01, F_{(1,60)} = 0.39$, $p = .54$ and no significant effect of amount of practice $V = 0.02$, $F_{(2,60)} = 0.60$, $p = .55$.

Table 44: Experiment 4. Mean median response time (and standard deviation) for correct trials on the category assignment task, as a function of acquisition schedule, block, and amount of practice.

	Block		
Acquisition Schedule	1	2	Average
	Low		
Blocked	3234.42 (1778.54)	2606.73 (1173.82)	2920.58 (1303.50)
Random	3519.59 (1862.65)	2391.86 (1076.61)	2955.73 (1336.29)
Average	3365.12 (1783.22)	2508.25 (1111.21)	2936.69 (1289.65)
	Medium		
Blocked	2702.75 (894.40)	2594.85 (1399.79)	2648.80 (1133.24)
Random	3490.92 (1203.05)	3044.54 (1273.83)	3267.73 (1162.47)
Average	3132.66 (1123.53)	2840.14 (1319.93)	2986.40 (1165.21)
	High		
Blocked	2725.68 (813.59)	2632.05 (1002.55)	2678.86 (903.24)
Random	2746.39 (929.12)	2333.17 (947.92)	2539.78 (914.57)
Average	2735.00 (843.78)	2497.55 (964.68)	2616.28 (886.91)
	Average		
Blocked	2913.46 (1278.59)	2611.43 (1157.69)	2762.44 (1109.69)
Random	3291.38 (1405.58)	2620.11 (1137.40)	2955.74 (1165.89)
Average	3096.69 $(1344.78)_a$	$2615.64(1139.05)_b$	2856.16 (1132.67)

Note. Means in each row that share a subscript do not differ significantly ($p < .05$).

Furthermore, there were no significant interactions between acquisition schedule and amount of practice, $V = 0.02$, $F_{(2,60)} = 0.63$, $p = .53$, between amount of practice and block, $V = 0.08$, $F_{(1,60)} = 2.71$, $p = .07$, nor between acquisition schedule and block, $V = 0.04$, $F_{(1,60)} = 2.42$, $p = .13$. Finally, there was no significant interaction between acquisition schedule, amount of practice, and block, $V < 0.01$, $F_{(2,60)} = 0.05$, $p = .95$.

5.2.3.4 Novel categories task

The novel categories transfer task had six blocks of 19 trials (16 positive and 3 negative example instances per block). Three of the blocks were ordered according to a blocked practice schedule and three blocks were ordered according to a random practice schedule. All participants completed both blocked and random retention blocks; the order of blocks (i.e., blocked first or random first) was counterbalanced across participants. Each participant's Median RT for correct, positive example instance trials was computed for each block. The blocked retention blocks were restructured so that each block had five trials of each of the three play categories; this required dropping the last trial of each block.

Figure 41 shows median RT for each block as a function of acquisition practice schedule and amount of practice. Both the blocked and random transfer blocks are displayed. Practice schedule and amount of practice did not reliably affect median RT in either the blocked or the random retention blocks.

5.3 Discussion

The acquisition data show main effects of practice schedule and amount of practice. A random schedule results in longer RT. Additionally, increasing the amount of practice results in shorter RTs for both practice schedule conditions. The acquisition curves show rapid improvement early, especially for the random conditions. When acquisition training is ended at four blocks (i.e., the low amount of practice) there is still more than a 500 ms difference in speed between practice schedule conditions. Additionally, with a low amount of practice there are large changes in error from block to block. Only by the last acquisition block (Block 4) do the errors for the random group drop below three, suggesting that participants are beginning to consistently discriminate between positive and negative examples. Moreover, the rule learning data support the error data: the low amount conditions have more difficulty explicitly verbalizing and drawing the rules that determine play category membership.

Retention and transfer measures show no differences between practice schedules and amount of practice, suggesting that the difference in performance at the end of

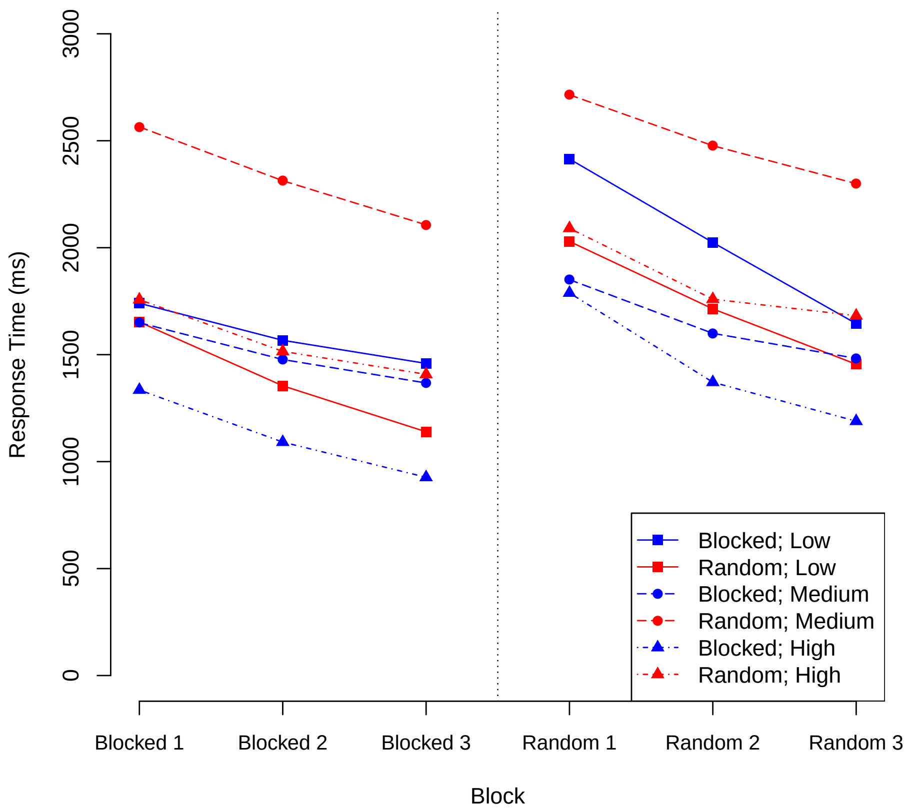

Figure 41: Response time for correct responses to trials with positive positive example instances during the novel categories transfer task, as a function of practice schedule and amount of practice.

acquisition training in Session 1 does not result in reliable differences in RT performance or errors after a 48-hour delay.

These data support the trends identified in Experiment 3. The categorization task seems to have two components underlying performance: rule learning and categorization proficiency. Rule learning is most prominent early in acquisition, and also the component that causes the biggest differences in performance between practice schedules. With continued acquisition training categorization proficiency becomes the stronger determinant of performance and the difference between schedules begins to disappear (though the random schedule condition still shows slower performance at the end of training). Retention and transfer do not show effects of acquisition practice schedule, suggesting that categorization proficiency, not rule learning, are the stronger determinants of performance.

This switch from rule learning to categorization proficiency over the course of skill acquisition is consistent with stage theories of skill acquisition. For instance, rule learning likely corresponds to the encoding stage and rapidly processing stimuli to determine category membership fits with the procedural stage in ACT-R (Neves & Anderson, 1981). Likewise, it is consistent with human abilities research suggesting that the determinants of task performance change over the course of skill acquisition (e.g., Fleishman & Hempel, 1955).

CHAPTER VI

GENERAL DISCUSSION

Practice schedule effects operate on multiple stages of information processing, with different mechanisms underlying performance at these different stages.

6.1 Practice Schedule Effects in Response-oriented Stages

In both experiments using a multisegment movement task, practice schedule has different effects on performance during acquisition, retention, and transfer. Over the entire course of acquisition training a random schedule consistently increases movement timing error (i.e., RMSE), relative to a blocked or blocked repeated practice schedule. The difference in error between schedules decreases with extended training, but is not eliminated. This suggests that even when performance has stopped improving rapidly, a context of changing task from trial to trial continues to add small amounts of additional error in response production.

On retention (and some transfer measures) a random acquisition schedule results in lower error than a blocked acquisition schedule. This effect appears to be reliable regardless of whether low, medium, or high amounts of acquisition training are used.

The pattern of results is consistent with a reconstruction view of contextual interference (Lee & Magill, 1983). Because the movement pattern changes on every (or every other) trial, a random schedule requires individuals to repeatedly retrieve and prepare a motor program[1]. These repeated retrievals increase participants' ability to explicitly recall the movement patterns and improves their ability to perform the

[1]In performing the multisegment movement task used in Experiments 1 and 2 individuals might use multiple motor programs for a given movement pattern. These motor programs might be joined together with continued practice. The nature of the explanation does not change, however, if individuals use multiple motor programs.

movement patterns. Moreover, participants trained with a random practice schedule are more adept at modifying these previously learned action plans to produce novel (mirror-reversed) movement patterns. This might arise because participants have more practice retrieving and loading motor programs into working memory and preparing those motor programs (c.f., Carlson & Shin, 1996), which could reduce the need to make changes to the motor programs during response execution, improving timing accuracy (Klapp & Wyatt, 1976).

The blocked-repeated schedule in Experiment 1 provides further support for the elaboration view. Error in the blocked-repeated conditions spikes each time the movement pattern switches. This is consistent with the theory that a motor program is retrieved from long-term memory, prepared, and then executed for a series of trials using the same movement pattern. The difference in retention performance between the blocked and blocked-repeated schedule conditions suggests that the process of retrieving and loading the motor program several times is enough to improve explicit recall and skill performance after a retention interval.

Figure 42 shows a model of task performance, adapted from Netick and Klapp (1994) and Keele (1968). The first process, response preparation, must occur each time a task switch occurs. In contrast, the second process, response execution, happens on each trial regardless of whether a task switch occurs. Based on the data showing increased error on all blocks (for the random schedule) and increased error on blocks following a task switch (for the blocked-repeated schedule), I argue that task switches elicit the response preparation process. Repeated response preparation practice facilitates learning, as demonstrated by the improved retention and transfer performance for the random group.

This model explains why, even with extended practice, a random schedule still produces higher error: the response preparation process is still elicited on these trials. Further, the model predicts that increasing the number of repetitions in the

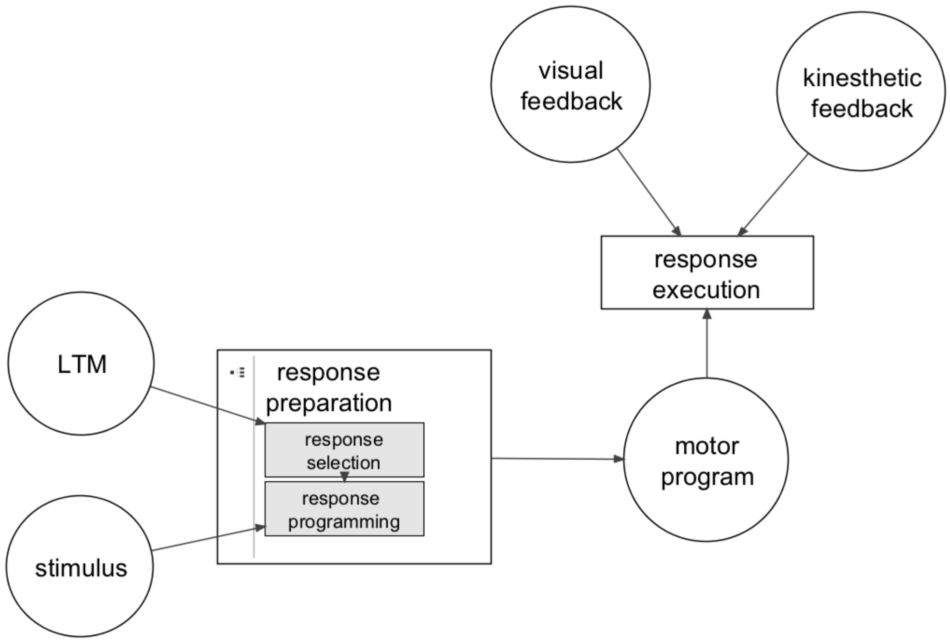

Figure 42: Hypothesized model of task performance in the multisegment movement task.

blocked-repeated schedule will produce monotonically increasing performance on retention and transfer measures. Additionally, it makes a novel prediction: one could create alternate manipulations to elicit this response preparation process, even within a blocked schedule. For instance, during the intertrial interval one could add an event to elicit response preparation without requiring response execution (e.g., asking participants to imagine performing a different motor movement). Performance in this *blocked schedule + imagining* condition is predicted to match performance of the random schedule, not the blocked schedule.

6.2 Practice Schedule Effects in Stimulus-oriented Stages

In both experiments using the visual categorization task, practice schedule has different effects on performance during acquisition, retention, and transfer. Early in acquisition training a random schedule increases response time, relative to a blocked or blocked-repeated practice schedule. However, with additional training the difference in RT between the random schedule condition and a blocked (or blocked-repeated) schedule condition disappears. This suggests that with enough acquisition training, a context of changing tasks from trial to trial does not significantly slow decision time.

On retention and transfer measures a random acquisition schedule does not cause differences in performance relative to training with a blocked or blocked-repeated acquisition schedule. This suggests that the context of changing the task from trial to trial does not result in better ability to categorize examples after a delay, or transfer the ability to novel categorization tasks.

The pattern of results suggests that performance on the categorization task is a function of two primary factors: rule learning and categorization decision speed. Rule learning (inductive learning) is affected by practice schedule: a random schedule increases categorization errors and contributes to longer decision times. However, this effect persists only as the rules are being learned. Once the rules have been learned then categorization decision speed seems to be the stronger determinant of response time. A random schedule seems to affect categorization decision speed during early acquisition trials (when rules are being learned or have recently been learned). In contrast, a random schedule does not seem to affect categorization decision speed during retention and transfer. A random schedule does seem to make this task more difficult primarily by making induction of the category rules more difficult.

This explanation of practice schedule effects on rule learning is consistent with research on inductive learning that suggests blocking examples makes the critical features more salient (Medin & Bettger, 1994) and improves schema abstraction (Elio & Anderson, 1981).

The blocked-repeated schedule in Experiment 3 provides further data to identify the locus of practice schedule effects on rule learning. Performance in the blocked-repeated schedule condition closely matches performance in the blocked schedule condition during acquisition. In the motor tasks (Experiment 1) there are performance

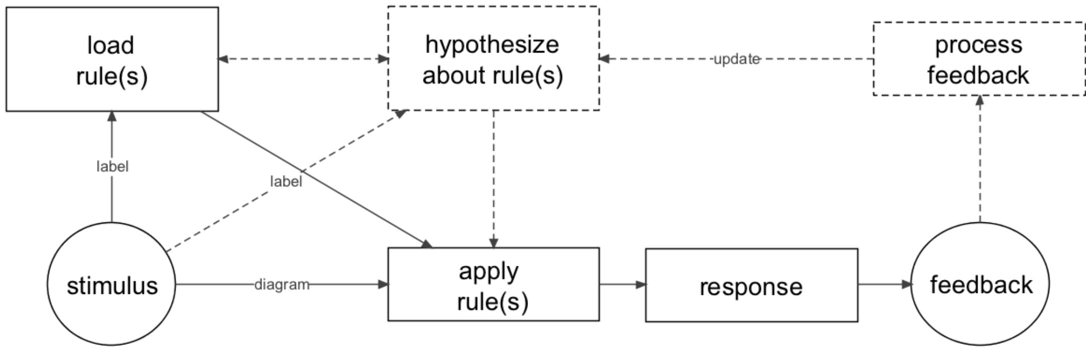

Figure 43: Hypothesized model of task performance in the categorization task. Dotted lines indicate processing that is present only during early stages of acquisition.

decrements following task switches; these performance decrements are absent in the categorization task (Experiment 3). The absence of performance decrements suggests that switching task categories does not increase categorization decision time.

Figure 43 shows a model of task performance based on these arguments. The dotted lines indicate processes involved in rule learning that are assumed to “drop out” once the rules are learned. Early in acquisition, processing mainly focuses on testing rules: hypothesizing about rules and processing the feedback can occur on all trials, regardless of whether the task switches. Experiments 3 and 4 suggest that this processing is aided by blocking multiple examples together on consecutive trials.

Later in acquisition, processing involves loading and applying the relevant, already-learned rules. On all trials an individual must apply the category rules to determine whether the diagram is a positive example instance of the category. However, the process of loading rules might be elicited only when the practice schedule contains task switches (i.e., each task switch requires rules to be loaded). When the play category switches from one trial to the next (e.g., on most trials in the random schedule and on the 12 task switch trials in a blocked-repeated schedule) then a new set of rules must be applied. If these rules are not already active then this set of rules might need to be loaded (i.e., retrieved from long-term memory).

If loading these rules occurred at all task switches (and only at task switches), and this loading was done prior to encoding the stimuli or applying the rules, then one would expect task switching to result in increased response time. Further, if loading these rules occurred at all task switches (and only at task switches), and this process of loading could be improved with repeated practice, then one would expect task switching to result in improved performance on retention or transfer trials.

Neither increased response time nor superior performance on retention or transfer trials was observed for the blocked-repeated or random schedules. This could be for several reasons:

- Rules for each play category can be held in WM simultaneously, eliminating the need to load (retrieve) them at each task switch.
- Rules might be able to be loaded concurrently with other stimuli processes, such as stimulus encoding or rule application. Only if the process of loading rules took a longer duration than the other parallel processing would one be able to see the effects of the rule loading on response time.
- Rules might be loaded at task switch, but this process is not difficult (i.e., does not show large improvements with practice) and does not add a significant amount to total processing time.

If any of these reasons are true, then one would not observe increases in response time or better retention and transfer performance as a function of practice schedule. The nature of the specific categorization task (e.g., number of category dimensions, number of rules, discriminability of features) might affect whether these reasons occur. This is an area for future research.

6.3 Stage Process Theory of Practice Schedule Effects

Data from Experiments 1–4 support the theory that practice schedules have different effects across different components of task performance. Figure 44 shows a hypothetical task that combines the processing from the tasks used in Experiments 1–4. Practice schedule affects the early stage of information processing (stimulus-oriented stages) differently than later stages of information processing (response-oriented stages). Response-oriented stages of information processing (as measured by the motor task) are affected by practice schedules. The contextual interference introduced by a random schedule results in more practice preparing the response, which improves retention and performance. Stimulus-oriented stages of information processing (as measured by the categorization task) are also affected by practice schedules, but the nature of this effect depends on the cognitive processing underlying the task. Learning which rules determine category membership is hindered by the contextual interference introduced by a random schedule. On the other hand, the present experiments provide no evidence that contextual interference affects the process of applying rules.

6.4 Amount of Practice and Practice Schedule Effects

6.4.1 Does amount of practice moderate practice schedule effects?

Amount of practice was investigated to determine whether it interacts with practice schedule, potentially moderating the effects of schedule on performance with different amounts of practice. Consistent with the stage process theory of practice schedule effects, I examined the role of amount of practice in moderating practice schedule effects at the response-oriented stages and the stimulus-oriented stages.

There is limited research on the role of amount of practice in moderating practice schedule effects in tasks that heavily weight response-oriented stages. Previous studies on amount of practice with practice schedule showed conflicting acquisition

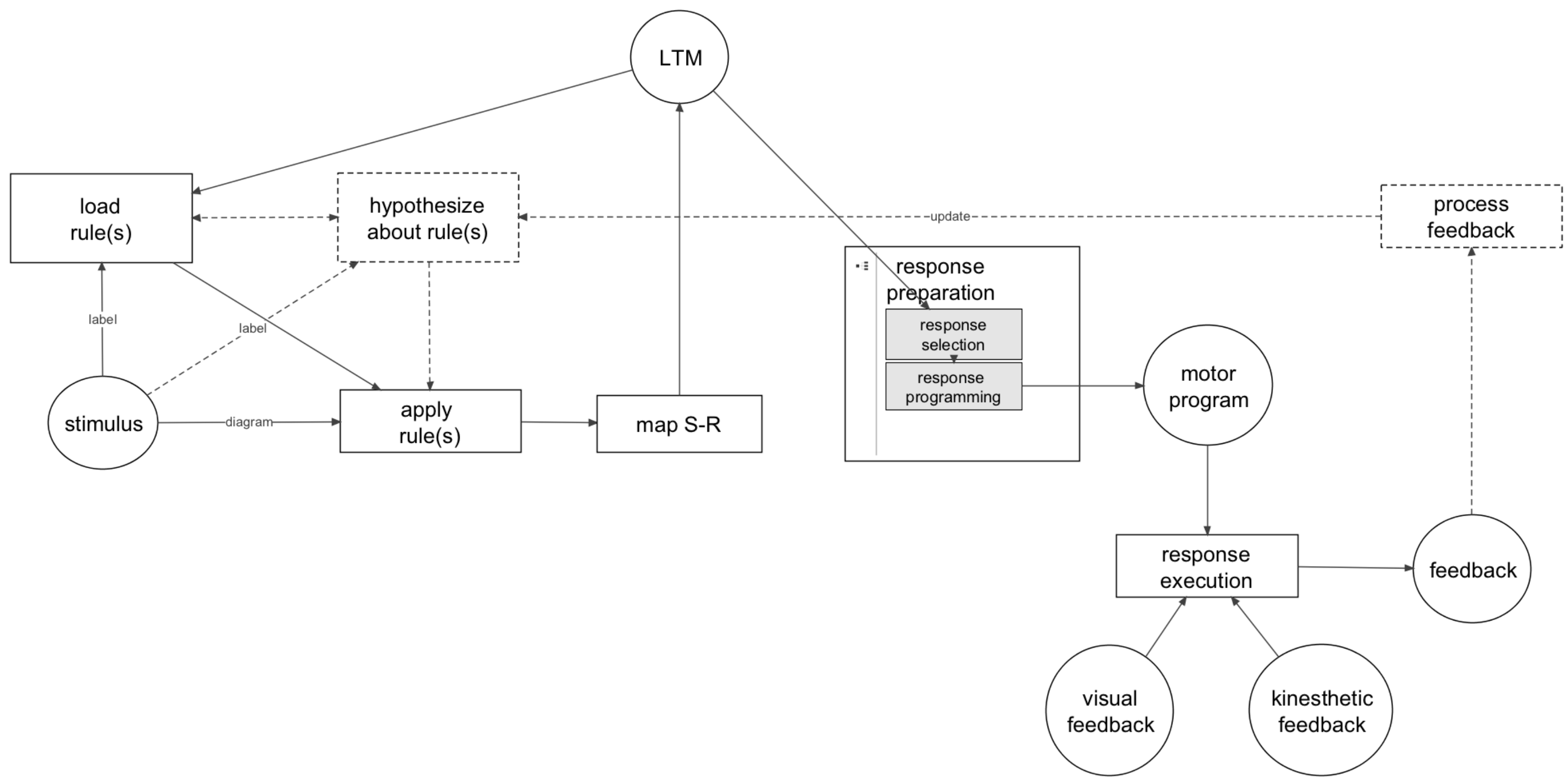

Figure 44: Hypothesized model of task performance in a hypothetical task that shares features with the categorization and motor task in Experiments 1 and 3. Dotted lines indicate processing that is present only during early stages of acquisition.

results in terms of whether the random schedule consistently added error throughout acquisition. Some evidence suggested that the learning curves did not converge during training (Giuffrida et al., 2002; C. H. Shea et al., 1990), while other evidence suggested the curves did converge (Proteau et al., 1994). Drawing conclusions from these data is difficult, however, due to varied dependent measures, different practice schedules, and lack of statistical tests to characterize performance changes across acquisition. By extending acquisition training for a longer duration, using RMSE (an overall measure of error), testing three practice schedules, and fitting learning curves to the data, data from Experiment 1 clarifies the literature. Learning curves for the blocked and random practice schedules did not converge by the end of training. Over the duration of acquisition training, practice schedule had consistent effects: a random schedule always had higher error than a blocked schedule. Only one post-acquisition measure, performance on the random retention block, showed a moderating effect for the extreme amounts of practice on practice schedule: practice schedule did not affect error when training stopped after a low amount of practice but did when training stopped after a high amount of practice. As these data indicate, there was no benefit for a blocked practice schedule over a random schedule. The empirical results showing a benefit of a blocked schedule with low amounts of training (C. H. Shea et al., 1990) has not been replicated in this study, nor in other research (Giuffrida et al., 2002; Proteau et al., 1994).

There is no research on the role of amount of practice in moderating practice schedule effects in tasks that heavily weight stimulus-oriented stages. The evidence collected in Experiments 3 and 4 suggests that amount of practice does moderate practice schedule effects. As seen in both the acquisition curves for errors and response time, the random schedule performs worse than the blocked group during early stages of acquisition, but not during later stages of acquisition. These practice schedule effects disappeared depending on the type of processing: early in acquisition (for rule learning processing, as evidenced by the drop in errors during early acquisition blocks) and later in acquisition (for categorization decisions, as evidenced by the convergence of RT curves during late acquisition blocks). No post-acquisition measures showed moderating effects of practice schedule. Knowledge of category rules was not affected by practice schedule for any of the levels of amount of practice. The acquisition error data, however, suggests that if *fewer* acquisition trials had been used in the low condition then the random schedule condition would have had lower knowledge of category rules, relative to the blocked schedule condition.

6.4.2 Defining amount of practice

One critique of the approach of studying amount of practice to moderate practice schedule effects is that the operationalization of amount of practice is not consistently applied (see the Critique of Existing Research section). One method to compare amount of practice is to define low and high amounts as the percentage of trials used by the medium amount of practice condition (see Table 4). In Experiments 2 and 4, low amount was defined as 25% and 50%, respectively, and high amount was defined as 200% for both. Note that Experiment 2 defined a low amount as the same percentage of medium trials as did C. H. Shea et al. (1990) but did not find the interaction effect that they showed.

As explained earlier, defining amounts of practice as a percentage of the number of trials in the medium condition is problematic because it is tied to the definition of the medium amount condition, which itself is arbitrary. Therefore, I developed an approach to quantify amount of practice based on learning curves fit to the task–practice schedule pairs. Rate of performance improvement (the first derivative of the learning curves; see Appendix A) is calculated for each task, as a function of the number of trials used in each amount of practice condition. Table 45 uses this approach

Table 45: Comparison of the rate of performance improvement for each amount of practice in Experiments 2 and 4.

	Amount		
Acquisition Schedule	Low	Medium	High
Motor Task			
Blocked	-0.679	-0.130	-0.057
Blocked-repeated	-0.486	-0.099	-0.044
Random	-0.939	-0.176	-0.076
Categorization Task			
Blocked	-2.928	-0.603	-0.274
Blocked-repeated	-4.593	-0.870	-0.379
Random	-17.539	-2.350	-0.860

to quantify the amounts of practice used in Experiments 2 and 4. The rate of performance improvement is much greater in the categorization task, relative to the motor task. In the categorization task there are two sub-tasks: learning the category rules and learning to making categorization decisions quickly. This dramatic differences in the estimates of rate of performance improvement between Experiments 2 and 4 likely arise, in part, because the rapid rule learning that occurs early in acquisition allows large reductions in time to categorize example instances.

6.5 Engineering Practice Schedules for Training Interventions

Introducing practice schedule manipulations into training interventions needs to be done carefully, with a solid understanding of the components of information processing that contribute to task performance as well as how different practice schedules affect processing in these components. If one were to manipulate practice schedules when training complex skills, multiple cognitive processes could be affected by the practice schedules. In this situation the context arising from a random schedule might or might not create interference, and the effects of that interference might vary across the components of task performance.

If one were interested in selecting a practice schedule to train a target skill, a part-task training paradigm (e.g., Frederiksen & White, 1989) might be needed to focus practice schedule effects on specific component-processes. This would require conducting a task analysis to identify the components of information processing underlying task performance. These components could then be trained individually, using part-task training. After identifying the component to be trained, the instructional designer could select a practice schedule that optimizes performance on the measure(s) of interest. Table 46 shows the optimal practice schedule for each outcome measure, as a function of the main task component that determines successful task performance.

If one is interested in training a category rule learning component, than a blocked or blocked-repeated schedule would require fewer trials for rule learning. This difference lasts only until the rules are learned, however. These blocked and blocked-repeated schedules do not seem to have any drawbacks in terms of declarative recall or retention and transfer performance.

If one is interested in training a categorization decision component, then a blocked or blocked-repeated schedule would result in quicker decision times. At some point during skill acquisition, this difference in decision time among schedules will disappear, but it requires more training trials than the category rule learning component. Additionally, this difference in decision time among practice schedules is likely exacerbated if the rules are not learned prior to training the category decisions.

If one is interested in training a response production component, then the choice of practice schedule depends on which outcome measure(s) the instructional designer is interested in optimizing. If the designer is interested in optimizing acquisition performance then a blocked schedule would be best. Alternately, a blocked-repeated schedule could be used, but would result in worse performance on those trials following a task switch; instructional designers (or trainers) should understand that these drops

Table 46: Matrix of practice schedules for optimizing performance on various outcome measures.

	Outcome measure			
Task focus	Acquisition	Declarative recall	Retention	Transfer
Category rule learning	Blocked, Blocked-repeated	Any	Any	Any
Categorization decision	Blocked, Blocked-repeated	Any	Any	Any
Response production	Blocked	Blocked-repeated, Random	Random	Random

in performance will occur on these trials. If the designer is interested in optimizing declarative recall, then a blocked-repeated or random schedule could be used. If the designer is interested in optimizing retention or transfer performance then a random schedule would be best.

After training individual components, training should focus on performing the whole task (Kirlik, Fisk, Walker, & Rothrock, 1998). When integrating these components, one again has to select a practice schedule. After the component skills have been trained, which practice schedule is most effective for integrating whole task performance? Answering this question requires a systematic approach to isolate and measure practice schedule effects. The proposed framework could be applied to training the whole-task integration also. For instance, if stimulus categorization and response production components of the task are trained individually, training would later need to focus on learning to associate the appropriate stimulus and response. This association might be best learned via a random schedule because it would create spaced practice on the associations (Melton, 1967). Understanding practice schedule effects in terms of their effects on component processing (both when the part task and whole task training is used) and their potential to be moderated by amount of practice would allow instructional designers to skillfully introduce practice schedule manipulations to optimize the training environment.

6.6 Limitations of the Present Studies

The present experiments do have several limitations. First, the tasks selected provide inevitable limitations on the generalizations of practice schedule effects. As argued by the theory, the effects of practice schedule depend on the components of information processing. In the present experiments tasks were selected such that they had minimal divergent information processing components (i.e., minimal stimulus-oriented processing in Experiments 1 and 2 and minimal response-oriented processing in Experiments 3 and 4). Tasks where performance depends on more equal weighting between component processing might have less clear effects. Additionally, task specific features might affect the results. For example, the four multisegment movement tasks used in Experiments 1 and 2 are hypothesized to use different motor programs; if the tasks were redesigned to use a single motor program (e.g., use one keypress sequence with four different goal times) then practice schedule effects might not be observed (Magill & Hall, 1990). Likewise, the play categories used in Experiments 3 and 4 are bidimensional categories with two features per category. Each feature requires minimal perceptual processing. If the rules or perceptual discriminations were more complicated there might have been more persistent differences in reaction time as a function of practice schedule.

A second limitation is that cleanly discriminating practice schedule effects at different stages of information processing relies on making these stages discrete. In the framework this is accomplished by selecting tasks that are representative of different stages of processing. However, as with Experiments 3 and 4, multiple stages (i.e., learning rules and making categorization decisions) might occur within that task. When they do, and if the stages are performed in parallel, then using RT as a measure of processing might not allow one to see differences between stages.

A third limitation is the learning curves used to fit the empirical data and calculate rate of performance improvement. Experiments 1 and 3 used learning curves based on fitting a simple power law to the data; the data fit well (except for the blocked-repeated data in Experiment 1), but their fit could be improved. Instead of aggregating participant data by schedule and fitting a power curve, individual participant data could be fit to exponential curves (Heathcote, Brown, & Mewhort, 2000). Further, model fit might be improved by modeling the performance on participant–task pairs (e.g., Participant 23 on Task A), which might show a series of different exponential curves and learning rates (Lacey, Krawitz, Kopecky, Kieras, & Meyer, 2004).

6.7 Future Directions

Future research should focus on mapping the space of categorization task features in order to systematically explore whether practice schedules have effects on other types of processing in stimulus-oriented stages, aside from rule learning. For instance, Kornell and Bjork (2008) argue that a random schedule helps in their task of learning to categorize paintings by painter style. Further, Gane and Catrambone (2010) present preliminary data showing practice schedule effects on response time in a task that involves categorizing poker hands. These conflicting results in the literature need to be further investigated. A clear mapping of task features and practice schedule effects, focused on the component processing that changes as a function of task switches, could help illuminate the features of categorization tasks that are influenced by practice schedule.

Extending the idea of using task switches, additional alternate practice schedules need to be explored. Blocked-repeated schedules can be created that vary the number of repetitions of the blocked schedule, which increases the number of task switches. These blocked-repeated schedule variations have the potential to further the understanding of practice schedule effects while suggesting beneficial schedules for training. Additionally, other schedules might exist in the space between blocked and random. For instance, Lee and Magill (1983) evaluated a serial schedule that switched tasks on each trial, but did so by using the same order of tasks across the entire schedule (see Figure 1).

As more empirical data are collected and theories of practice schedules become better specified, investigators should research how practice schedules can be mixed together for individual participants. For instance, one common, intuitive idea is to use a blocked schedule early in training and transition to a random schedule later in training. This method needs to be empirically tested; there is a danger that processing strategies learned early in acquisition with a blocked schedule might limit processing that could occur in later acquisition with a random schedule (e.g., Doane et al., 1996). Further, when should the practice schedule switch? The method I developed to quantify amount of practice (i.e., estimate the rate of performance improvement; Appendix A) could be used to identify potential points at which the schedule should be switched.

These efforts to specify the task features, characterize alternate practice schedules, and to design optimal mixes of practice schedules have the potential to facilitate the development and implementation of automated training programs. Computerized training tools have begun to use information from decades of research on spacing effects (e.g., Balota, Ducheck, Sergent-Marshall, & Roediger, 2006; Hintzman, 1974) to select "optimal" intervals to present the to-be-learned item. As training for more complex tasks with multiple task variants is developed, traditional models of spacing effects will no longer be adequate. A strong theoretical model of practice schedule effects, how these effects occur across task types and varying levels of practice, is needed to provide instructional designers with the knowledge to create optimal practice schedules for task training.

APPENDIX A

DETERMINING RATE OF PERFORMANCE IMPROVEMENT

I propose a novel method for defining amount of practice. This method involves fitting learning curves to empirical data, calculating the rate of performance improvement function of these curves, and then using that rate to determine the number of trials to define each amount of practice. I used acquisition data from the three amounts of practice studies (Giuffrida et al., 2002; Proteau et al., 1994; C. H. Shea et al., 1990) to determine the best estimate for these criteria values.

To begin, I created learning curves and rate of performance improvement curves. First, I used the published figures in each study to estimate point values for the RT data at each block; this allowed me to calculate learning curves for each practice schedule. I fit each data set to a simple power curve[1]

$$Y = B + N^{-\alpha} \tag{3}$$

where Y is performance, N is trial number and B is initial performance (Newell &

[1] I used the simple power law formula and therefore fit a power function. Other functions (e.g., an exponential or the general power law) could be used instead, and the same procedure of calculating the first derivative and then selecting an optimal criterion would apply. The simple power curve does the best at satisfying my goals for the curve. There is some concern that a power function only fits better when data is aggregated, and that individual data is better fit by an exponential function (e.g., Heathcote et al., 2000). However, I am using aggregated data to determine the number of trials for each amount of practice, and so I am interested in a curve that fits aggregate data.

Additionally, one could use the general power law $Y = A + B(N + E)^{-\alpha}$ where A is the asymptote and E is "the number of trials of learning that occurred prior to the first trial as measured (i.e., prior experience)" (Newell & Rosenbloom, 1981, p. 19). However, as the equation is underidentified, solving for A and E requires a search through the possible space of A and E values and evaluating the curve fit (e.g., R^2) of each equation to determine the appropriate values. Newell and Rosenbloom acknowledge this and offer a solution: "The difficulty of course is that A and E are not known in advance, so the curve cannot be plotted as an initial exploratory step in an investigation. One alternative is just to plot in $\log(T) - \log(N)$ space and understand the deviations" (p. 19). I propose the latter approach.

Rosenbloom, 1981). Ordinal Least Squares (OLS) regression with the log-transformed data provided the learning rate parameter (α) and initial performance parameter (B). The rate of performance improvement function was obtained by computing the first derivative of the learning function

$$\frac{dY}{dN} = -\alpha B N^{-\alpha-1}. \tag{4}$$

This function ($\frac{dY}{dN}$) specifies how the rate of performance improvement changes as a function of the number of acquisition trials (i.e., velocity of the learning curve). Any point along this function describes the instantaneous rate of performance improvement (i.e., C) for a given number of trials. These learning curves and rate of performance improvement functions were calculated for each data set; Figures 45 and 46 display each. Some of the graphed functions use interpolated data to estimate early or late practice data.

A.1 Determining the Number of Trials

One can use these three C values to determine the number of trials that would define each amount of practice variable (i.e., number of trials). For example, in Figure 46 (C. H. Shea et al., 1990 data) $\frac{dY}{dN} = C_{medium}$ after 137 trials. Thus, on average, 137 trials are needed before the rate of performance improvement slows to -0.001, thus marking the end of the medium amount of practice. In contrast, in Figure 46 (Proteau et al., 1994 data) $\frac{dY}{dN} = C_{medium}$ after 162 trials. Thus, on average, 162 trials are needed before the rate of performance improvement slows to -0.001, thus marking the end of the medium amount of practice.

A.2 Testing Different Operationalizations of Amount of Practice

Each study has a different number of trials but because the same C values are used for all studies, the ratio of the number of trials making up each amount of practice

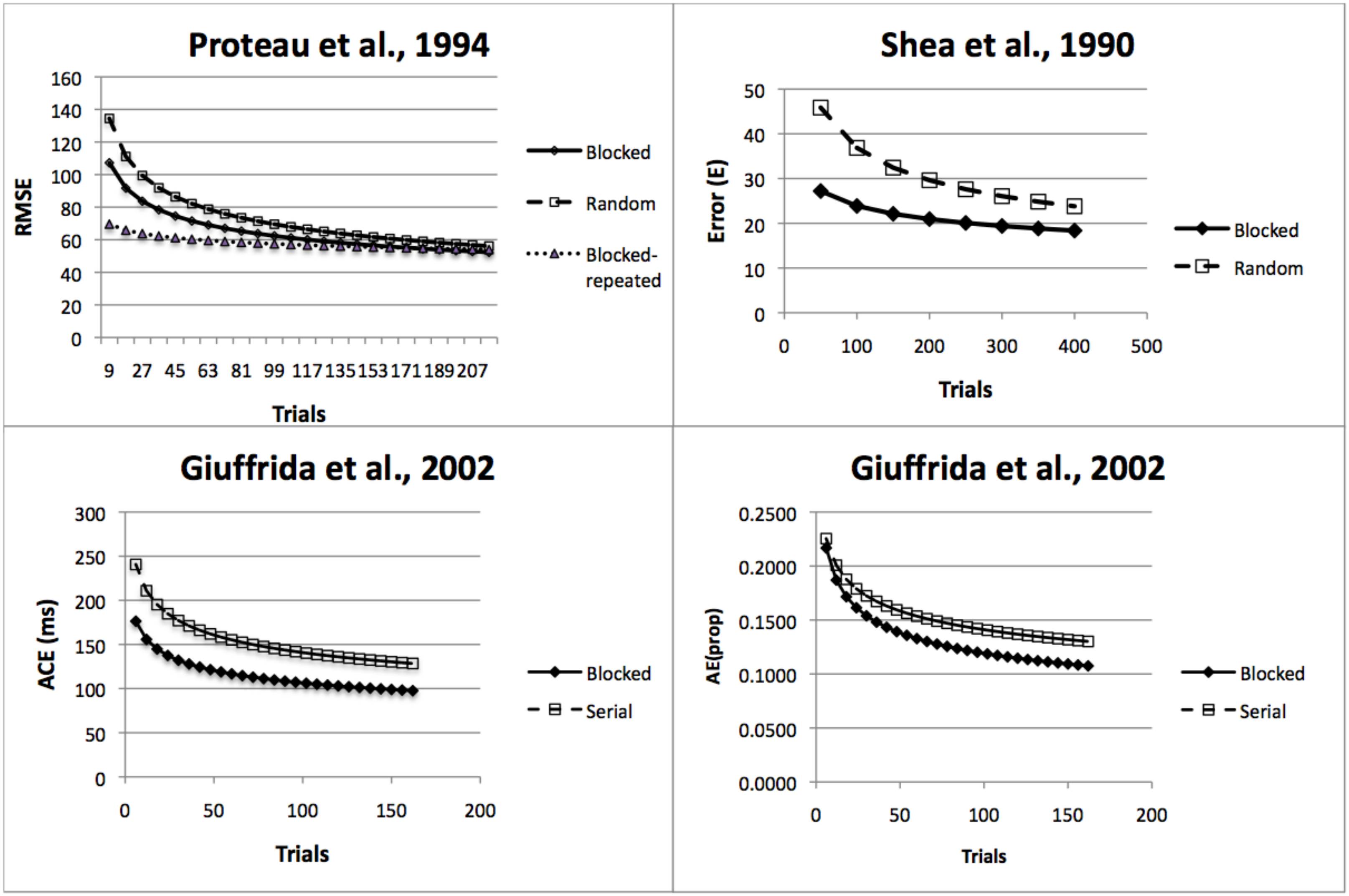

Figure 45: Learning curves generated for the four existing data sets.

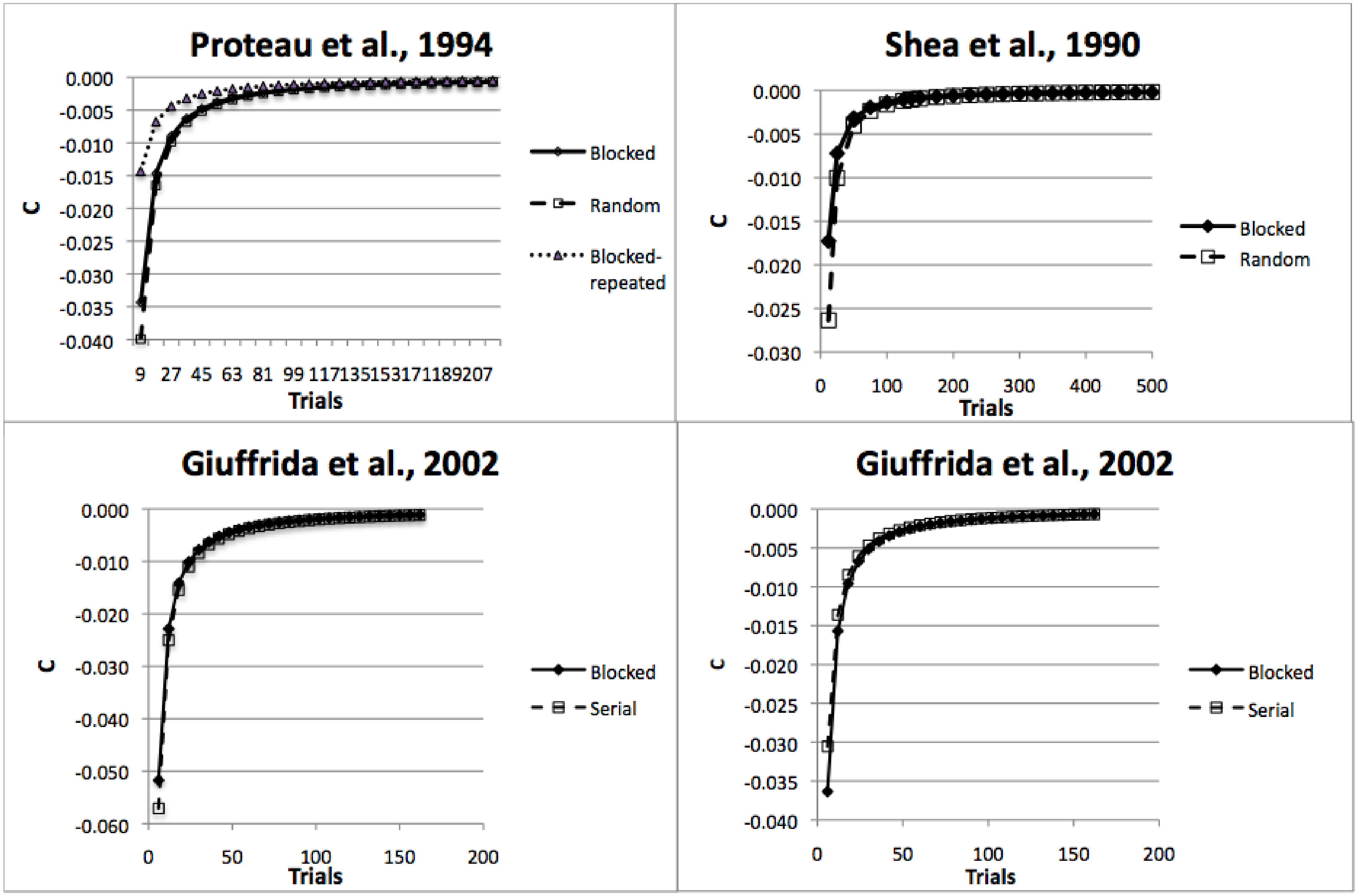

Figure 46: Rate of performance improvement curves generated for the four existing data sets.

condition are roughly equivalent. That is, low (C = -0.004) has about .33 of the trials as the medium (C = -0.01) and the high (C = -0.00025) has about three times the number of trials as the medium. In essence, C for the medium amount sets the number of trials for that task and, in general, the number of trials in the low and high amounts roughly follow a 0.33–1.00–3.00 ratio. Future investigations into amount of practice can test different definitions of amounts of practice using one of two methods: (1) increase or decrease the medium C value (e.g., C_{medium} = -0.002 or C_{medium} = -0.0005) but keep the 0.33–1.00–3.00 ratio (i.e., shift the set of definitions left or right), or (2) keep the medium C value and modify the ratio (e.g., 0.25–1.00–2.00). This allows consistent, task independent changes to the operational definition of amount of practice, but maintains a uniform measure to facilitate comparing these changes. It is possible, therefore, to conduct a series of studies using a range of C values and ratios to map the resulting acquisition and retention performance. Mapping this full space might allow one to define "optimal" values of practice.

These existing estimates of C (and thus the number of trials) also provides a new way to analyze the existing data. For instance, it suggests that both the Giuffrida et al. (2002) and Proteau et al. (1994) data should have *increased* the number of practice trials that participants completed in both the medium and high groups. This lends support to my critique of the existing studies that argued that these two studies also had the more complex tasks and thus might have needed longer acquisition practice.

APPENDIX B

DEPENDENT MEASURES IN EXPERIMENTS 1 AND 2

Root Mean Square Error (RMSE) will be the primary dependent measure. Conceptually, RMSE reflects both bias and inconsistency of responses (i.e., timing)

$$E^2 = VE^2 + CE^2 \tag{5}$$

where VE^2 is variable error (inconsistency of responding) and CE^2 is constant error (bias; deviation relative to target). Both VE^2 and CE^2 can be calculated separately.

Schmidt and Lee (2005) state that CE (or $|CE|$) will quickly decrease with practice. Therefore most improvement with practice comes from decreasing VE. This suggests that with extended practice VE might show the most change. However, the literature uses RMSE (Proteau et al., 1994; C. H. Shea et al., 1990) as a whole, and the one study that did analyze CE (Giuffrida et al., 2002) found that $|CE|$ decreased more for the blocked schedule condition than the serial schedule condition and that differences in |CE| between practice schedules remain at the end of acquisition. Thus, it seems RMSE is good primary measure, especially if one is interested in characterizing performance with different amounts of practice.

APPENDIX C

EXPERIMENT 3 AND 4 DATA EXCLUSION CRITERIA

Experiments 3 and 4 included catch trials (negative example instances) that were intended to encourage participants to complete the task as intended (i.e., process the stimuli and make an informed decision before responding). These catch trials were added primarily to reduce the possibility that participants in the blocked schedule condition would not process the stimuli because the same category of stimuli is shown repeatedly.

Data from the catch trials were also used to identify participants that adopted this type of strategy. I used these catch trials to define a data-driven exclusion criteria. The criteria was scored as follows. First, I counted the number of blocks in which both (1) all the catch trials were incorrect and (2) *only* the catch trials were incorrect. Second, I eliminated any of these blocks that had a median response time (on both positive and negative example instances) greater than or equal to 900 ms. Therefore, this joint criteria is that within a block all negative trials (and only negative trials) were missed, while responses on all trials in that block took less than 900 ms. If 25% or more of a participant's acquisition blocks met this criteria then the participant was excluded. This exclusion criteria corresponds to a participant pressing "d" (the response for legitimate examples, which is therefore the default response) for all 19 trials in a block, with the median time per trial under 900 ms on 25% or more of their training blocks. In other words, on at least one-fourth of the training blocks participants pressed the default response button at a rate that suggests they were not attending to the stimuli.

Unexpectedly, in both Experiments 3 and 4 participants in the random conditions were the only participants that met this criteria. Three participants in each experiment met this criteria.

APPENDIX D

FOOTBALL KNOWLEDGE PRETEST QUESTIONNAIRE

1. What does it mean to read a defense?
2. How many defensive players can be on the field?
3. Name the defensive player positions.
4. What is the difference between a defensive front and a defensive secondary?
5. Name up to five defensive formations.
6. What is a nickel defense?

APPENDIX E

CREATING PRACTICE SCHEDULES IN EXPERIMENT 3

In this section I describe the process of creating variations of each practice schedule (e.g., Blocked 1, Blocked 2, Blocked 3, etc.), and the consistencies across practice schedules (i.e., Blocked, Blocked-repeated, Random).

E.1 Acquisition

E.1.1 Distributing the location of negative example instances within blocks

Three of the 19 trials (16 %) in each block were negative example instances. Negative examples were randomly distributed within a block with two constraints: 1) the first two trials in a block were always positive examples and 2) the last trial in a block was always a positive example. The negative trials were randomly distributed for each of the 24 acquisition blocks. The same distribution of negative example trials within acquisition blocks was used for all 10 acquisition schedules. Thus the distribution of positive and negative examples throughout acquisition trials was constant across practice schedule.

E.1.2 Generating example instances

For each category (e.g., blue) the specific example instances were created based on the random distortion procedure described in the Experiment 3 Method section. This set of 384 positive example instances (96 instances per category) and 72 negative example instances (18 instances per category) was created once and then used for all 10 schedules. The order of these example instances were randomized separately (i.e., positive randomized, negative randomized) and then used for all practice schedules.

E.1.3 Positive examples

A trial has three components: *label*, *example instance: category*, *example instance: diagram*. If the label and category match it is a positive example; if they do not match it is a negative example.

E.1.3.1 Categories

In blocked schedules the order of play categories was determined by a Latin Square (see Table26). For example, the Blocked 1 schedule had six blocks of green followed by six blocks of blue, six blocks of gray, and six blocks of red. Blocked-repeated schedules used the same order of play categories as the corresponding blocked schedule (i.e., the Latin Square specified the order of both the blocked and blocked-repeated schedules). Random schedules were created by randomizing the order of play categories with the constraint that no more than two consecutive trials used the same play category. The Random 1 and Random 2 schedules used two different randomized orders.

E.1.3.2 Instances (diagrams)

Within each play category, the order of these example instances was consistent for all 10 schedules. The same order of example instances within a category was used for all blocked schedules. For example, the first positive example of the green category was named "green_160". Regardless of when the green blocks occurred (e.g., Block 1 in the Blocked 1 schedule but Block 13 in the Blocked 2 schedule), the first green trial always used the "green_160" diagram.

Additionally, *within a category* the order of example instances for the blocked-repeated schedules was consistent with all the blocked schedules. For example, the first example instance in the green blocks (e.g., Block 1 in the Blocked-repeated 1 schedule and Block 5 in the Blocked-repeated 2 schedule) always used the "green_160" diagram.

As with the blocked and blocked-repeated schedules, *within a category* the same order of example instances was used for all random schedules.

E.1.3.3 Labels

Because they are positive examples, for all schedules the label matched the example instance category.

E.1.3.4 Implications

At the category level (e.g., blue), the specific positive example instances, and the order of those instances, was constant across all blocked, blocked-repeated, and random schedules. In terms of positive examples, the only difference between schedules was the order of play categories.

E.1.4 Negative examples

E.1.4.1 Labels

In both the blocked and blocked-repeated schedules the label for the negative instances matched the label of the positive example instances for that block (e.g., both positive and negative example instances used the label Green for all trials in the block).

In the random schedules the label for the negative instances was randomly selected with the constraint that no more than two consecutive labels (positive or negative trials) used the same label. Because there were three example instances per block and four category labels to choose from, three of the four labels were randomly selected (without replacement) from the pool of labels and assigned to the negative instances for that block. For the next block any remaining labels in the pool were randomly selected and assigned to the negative instances for that block. The state of the pool was preserved across blocks and was repopulated when empty. Random 1 and Random 2 used two different orders of labels.

E.1.4.2 Categories

In the blocked schedule the order of negative example instance categories was determined by random selection. Because the positive example instances were all of one category, the remaining three categories were used for the negative example instances. Within each block the order of these categories was randomized. Each block's randomized order was used in all versions of the Blocked and Blocked-repeated schedules. Thus the order of negative example categories within a block was constant at the level of category blocks, but not at schedule (e.g., the first Green block had negative categories in the order gray, red, and blue. This first green block occurred on Block 1 in the Blocked 1 schedule and Block 13 in the Blocked 2 schedule.

The order of categories was constant for both Random schedules, but was different than the Blocked and Blocked-repeated schedules. The order of categories can not be the same as Blocked and Blocked-repeated because the Random schedule is sampling from a pool of four categories whereas Blocked and Blocked-repeated are sampling from a pool of three categories.

In the random schedule the category for the negative instances was randomly selected with the constraint that it did not match the label that had already been selected for that trial (because if it matched it would not be a negative example). Because there were three trials and four categories, I again used a pool of four categories, sampling (without replacement) and preserving its state across blocks.

The two random schedules did not match one another in terms of the order of negative example labels nor the order of negative example categories. Moreover, because the blocked and blocked-repeated orders necessitated a constraint on which negative examples could be used (i.e., it could not match the positive examples in the block), the random schedules did not match the blocked or blocked-repeated schedules in terms of the negative example labels nor negative example categories.

E.1.4.3 Instances

The same order of negative example instances within a category was used for all blocked schedules. For example, the negative example instance used for the first green negative example in the Blocked 1 schedule (i.e., Block 7, Trial 12) was also used as the first green negative example in the Blocked 2 schedule (i.e., Block 1, Trial 9), the Blocked 3 schedule (i.e., Block 1, Trial 9), and the Blocked 4 schedule (i.e., Block 19, Trial 12).

Blocked-repeated schedules used the same selection and order of example instances as the corresponding block in the respective blocked schedule. Extending the example from the previous paragraph, the negative example instances in the first Green block is the same in the Blocked 1 schedule as the Blocked-repeated 1 schedule (and also the same as the first Green block in the Blocked-repeated 2, Blocked-repeated 3, and Blocked-repeated 4 schedules).

The random schedules did match one another (and the blocked and blocked-repeated schedules) in terms of the order of negative example instances *within a category*. For example, the same negative example instance was used for the first green negative example trial (i.e., "green_122") across all 10 schedules. Additionally, the distribution of negative example instances within the blocks matched all the other schedules.

E.1.4.4 Implications

The distribution of negative example instances and the order of negative example instances (within a category) were constant across schedules; the order of negative example instance categories (and their labels) differed between the random schedules and between the random schedules and the blocked and blocked-repeated schedules.

E.2 Retention Task

In retention all participants complete one block of trials ordered in a blocked schedule and one block of trials ordered in a random schedule. The order in which participants completed the blocked and random orders was counterbalanced across participants (i.e., Blocked first and Random first).

E.2.1 Distributing Negative Examples Throughout Blocks

Following the procedure in acquisition, three negative example trials were randomly distributed in each block, with the same constraints as in acquisition.

E.2.2 Positive Examples

As in acquisition a set of positive and negative example instances was generated (based on the random distortion procedure); all participants saw the same set of positive and negative example instances.

In the blocked schedule, one order of play categories was used: red, green, gray, blue. This order was randomly generated with the constraint that the order could not match any of the blocked orders during acquisition (i.e., it was different from Blocked 1 – 4).

In the random schedule, one order of play categories was used. The random order was created following the same constraints used during acquisition.

E.2.3 Negative Examples

Labels for the negative examples were selected in different ways, on the basis of whether the order was blocked or random. In the blocked order the negative example labels matched their surrounding positive example labels; this preserves the feeling of a blocked structure. In the random order the three negative example labels were randomly selected from the pool of four categories.

The play categories for the negative example instances were selected in different ways, on the basis of whether the block was ordered according to a blocked or random schedule. In the blocked schedule block the first two trials were constrained to be gray and blue and the last trial was constrained to be red or green; the resulting random order was gray, blue, green. These constraints were necessary so that both the blocked schedule blocks could use the same play categories for the negative examples, regardless of whether it was used as the first or second block. In the random schedule block three negative example categories were randomly selected from the pool of four categories, with the constraint that the category could not match the label for that trial. The same random order (red, gray, green) was used for both random schedules, regardless of whether it was used as the first or second block.

The specific diagrams used for the positive and negative example instances were presented in the same sequence for both the blocked first and the random first conditions. In other words, regardless of whether participants completed a blocked or random block first, the orders of example instances *within the category* were the same.

E.3 Category Assignment Task

One schedule was used for the category assignment task. All example instances were randomly ordered following the constraint that no more than two consecutive example instances occurred. In this transfer task there was no category label, only the diagrams. Thus example instances were not split into negative and positive example instances. The set of example instances was created using the same random distortion procedure. Example instances were novel: they did not duplicate example instances used during acquisition.

E.4 Novel Categories Task

Creating the schedules for the novel categories task required a combination of methods used for creating the acquisition and retention schedules. In the novel categories task all participants completed three blocks ordered according to a blocked schedule and three blocks ordered according to a random schedule. The order in which they completed these blocks was counterbalanced across participants (i.e., blocked first and random first).

E.4.1 Distributing negative example instances throughout blocks

As with acquisition, three of the 19 trials (16%) in each block were negative example instances. Negative examples were randomly distributed within a block with the same two constraints as used during acquisition. The negative trials were randomly distributed for each of the six blocks. The same distribution of negative example trials within acquisition blocks was used for both orders. Thus the distribution of positive and negative examples throughout the transfer trials was constant across orders.

E.4.2 Play labels

E.4.2.1 Blocked schedule blocks

For the blocked blocks, the three play categories (cyan, purple, and yellow) were randomized to create one blocked order: purple, yellow, and cyan. This blocked order determined the play labels for both the positive and negative example instances within each of the three blocked blocks.

E.4.2.2 Random schedule blocks

The three play categories were randomized to create one random order of positive example instances (which spanned three blocks). This randomization followed the same constraint used to create the random schedules in acquisition. Because there were three categories and 16 positive example trials, each category was used at least five times each block. The 16th trial used one of the three categories, randomly selected without replacement from a pool of three.

For the random blocks, the three play categories were also randomized to create labels for the negative instances; each category was used as a label once per block.

E.4.3 Play diagrams

In both the blocked and random blocks the play categories for the positive example instances matched the play labels. The play categories for the negative example instances differed based on schedule. For the blocked schedule a pool of six categories (two repeats per category) was randomly selected without replacement and inserted into negative trials with the constraint that the category did not match the label. For the random order a pool of three categories was randomly selected (without replacement) and inserted into negative trials with the same constraint that the category did not match the label. The pool was repopulated when empty.

REFERENCES

Balota, D. A., Ducheck, J. M., Sergent-Marshall, S. D., & Roediger, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? explorations of spacing effects in healthy aging and early stage alzheimer's disease. *Psychology and Aging*, *21*, 19-31.

Battig, W. F. (1972). Intratask interference as a source of facilitation in transfer and retention. In R. F. Thompson & J. F. Voss (Eds.), *Topics in learning and performance* (p. 131-159). New York: Academic Press.

Battig, W. F. (1979). The flexibility of human memory. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (p. 23-44). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (p. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of william k. estes* (Vol. 2, p. 35-67). Hillsdale, NJ: Erlbaum.

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*(6), 546-556.

Carlson, R. A., & Shin, J. C. (1996). Practice schedules and subgoal instantiation in cascaded problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 157-168.

Carlson, R. A., & Yaure, R. G. (1990). Practice schedules and the use of component skills in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 484-496.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354-380.

Charney, D. H., & Reder, L. M. (1986). Designing interactive tutorials for computer users. *Human-Computer Interaction*, *2*, 297-317.

Doane, S. M., Alderton, D. L., Sohn, Y. W., & Pellegrino, J. W. (1996). Acquisition and transfer of skilled performance: Are visual discrimination skills stimulus specific? *Journal of Experimental Psychology: Human Perception and Performance*, *22*(5), 1218-1248.

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on practice. *Journal of Applied Psychology*, *77*(5), 615-622.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 397-417.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (p. 243-285). New York: Academic Press.

Fitts, P. M., & Posner, M. I. (1967). *Human performance.* Belmont, CA: Wadsworth.

Fleishman, E., & Hempel, W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology*, *49*, 301-312.

Fox, J., & Weisberg, S. (2011). *An r companion to applied regression* (2nd ed.). Washington DC: SAGE Publications, Inc.

Frederiksen, J. R., & White, B. Y. (1989). An approach to training based on principled task decomposition. *Acta Psychologica*, *71*, 89-146.

Gane, B. D., & Catrambone, R. (2009, April). *Practice schedules in cognitive skill acquisition: Effects of example order on categorization and problem solving.* Paper presented at the American Educational Research Association annual meeting, San Diego, CA.

Gane, B. D., & Catrambone, R. (2010). Order effects in categorization: Identifying "the nuts" in poker [abstract]. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (p. 713). Austin, TX: Cognitive Science Society.

Giuffrida, C. G., Shea, J. B., & Fairbrother, J. T. (2002). Differential transfer benefits of increased practice for constant, blocked, and serial practice schedules. *Journal of Motor Behavior*, *34*(4), 353-365.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185-207.

Hiew, C. C. (1977). Sequence effects in rule learning and conceptual generalization. *American Journal of Psychology*, *90*(2), 207-218.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The loyola symposium* (p. 77-97). Potomac, MD: Erlbaum.

Immink, M. A., & Wright, D. L. (1998). Contextual interference: A response planning account. *Quarterly Journal of Experimental Psychology*, *51A*, 735-754.

Jamieson, B. A., & Rogers, W. A. (2000). Age-related effects of blocked and random practice schedules on learning a new technology. *Journal of Gerontology: Psychological Sciences*, *55B*, 343-353.

Johnson, S. L. (1980). A cost analysis method of establishing training criteria. *Ergonomics*, *23*(12), 1137-1145.

Keele, S. W. (1968). Movement control in skilled motor performance. *Psychological Bulletin*, *70*, 387-403.

Kirlik, A., Fisk, A. D., Walker, N., & Rothrock, L. (1998). Feedback augmentation and part-task practice in training dynamic decision-making skills. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (p. 91-113). Washington DC: American Psychological Association.

Kirlik, A., Walker, N., Fisk, A. D., & Nagel, K. (1996). Supporting perception in the service of dynamic decision making. *Human Factors*, *38*(2), 288-299.

Klapp, S. T., & Wyatt, E. P. (1976). Motor programming within a sequence of responses. *Journal of Motor Behavior*, *8*, 19-26.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the"enemy of induction"? *Psychological Science*, *19*, 585-592.

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*(4), 239-243.

Lacey, S. C., Krawitz, A., Kopecky, J. J., Kieras, D. E., & Meyer, D. E. (2004, November). *Routine procedural recipes for rapid learning in choice-reaction tasks.* Paper presented at the meeting of the Psychonomic Society, Minneapolis, MN.

Lee, T. D., & Magill, R. A. (1983). The locus of contextual interference in motor-skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 730-746.

Lee, T. D., Wulf, G., & Schmidt, R. A. (1992). Contextual interference in motor learning: Dissociated effects due to the nature of task variations. *Quarterly Journal of Experimental Psychology*, *44A*, 627-644.

Li, Y., & Wright, D. L. (2000). An assessment of the attention demands during random- and blocked-practice schedules. *Quarterly Journal of Experimental Psychology*, *53A*, 591-606.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science*, *9*, 241-289.

McGuire, W. J. (1961). A multiprocess model for paired-associate learning. *Journal of Experimental Psychology*, *62*(4), 335-347.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin and Review*, *1*(2), 250-254.

Melton, A. W. (1967). Repetition and retrieval from memory. *Science*, *158*, 532.

Nagel, K. (1993). *Training pattern recognition: Using worked examples to aid schema acquisition.* Unpublished master's thesis, Georgia Institute of Technology.

Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, *140*(1), 48-77.

Netick, A., & Klapp, S. T. (1994). Hesitations in manual tracking: A single-channel limit in response programming. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 766-782.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (p. 57-84). Hillsdale, NJ: Erlbaum.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (p. 1-55). Hillsdale, NJ: Erlbaum.

O'Brien, R. G., & Kaiser, M. K. (1985). Manova method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316-333.

Ollis, S., Button, C., & Fairweather, M. (2005). The influence of professional expertise and task complexity upon the potency of the contextual interference effect. *Acta Psychologica*, *118*, 229-244.

Pellegrino, J. W. (1972). Effects of intralist response formal similarity upon paired-associate transfer and retroactive inhibition. *Journal of Experimental Psychology*, *92*, 134-142.

Posner, M. I., Goldsmith, R., & Welton, K. E., Jr. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, *73*(1), 28-38.

Proteau, L., Blandin, Y., Alain, C., & Dorion, A. (1994). The effects of the amount and variability of practice on the learning of a multi-segmented motor task. *Acta Psychologica*, *85*(1), 61-74.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*, 144-161.

Richland, L. E., Finley, J. R., & Bjork, R. A. (2004). Differentiating the contextual interference effect from the spacing effect. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the cognitive science society* (p. 1624). Mahwah, NJ: Erlbaum.

Ritter, F. E., Nerb, J., Lehtinen, E., & O'Shea, T. (Eds.). (2007). *In order to learn: How the sequence of topics influences learning.* New York: Oxford University Press.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207-217.

Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis* (4th ed.). Champaign, IL: Human Kinetics.

Schneider, V. I., Healy, A. F., & Bourne Jr., L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419-440.

Schneider, W. (1985). Training high-performance skills: Fallacies and guidelines. *Human Factors*, *27*, 285-300.

Shea, C. H., Kohl, R., & Indermill, C. (1990). Contextual interference: Contributions of practice. *Acta Psychologica*, *73*, 145-157.

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquistion, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 179-187.

Shea, J. B., & Zimny, S. T. (1983). Context effects in memory and learning movement information. In R. A. Magill (Ed.), *Memory and control of action* (p. 345-366). New York: North-Holland Publishing Company.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 907-912.

Simon, D. A., & Bjork, R. A. (2002). Models of performance in learning multi-segment movement tasks: Consequences for acquisition, retention and judgments of learning. *Journal of Experimental Psychology: Applied*, *8*, 222-232.

Smith, E. E., Chase, W. G., & Smith, P. G. (1973). Stimulus and response repetition effects in retrieval from short-term memory: Trace decay and memory search. *Journal of Experimental Psychology*, *98*, 413-422.

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General*, *134*(4), 443-460.

Stokes, P. D., Lai, B., Holtz, D., Rigsbee, E., & Cherrick, D. (2008). Effects of practice on variability, effects of variability on transfer. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 640 - 659.

Underwood, B. J. (1952). An orientation for research on thinking. *Psychological Review*, *59*, 209-220.

Underwood, B. J., & Schulz, R. W. (1960). *Meaningfulness and verbal learning*. New York: J. B. Lippincott Company.

van Gerven, P. W. M., Paas, F., van Merriënboer, J. J. G., & Schmidt, H. G. (2006). Modality and variability as factors in training the elderly. *Applied Cognitive Psychology*, *20*, 311-320.

van Merriënboer, J. J. G., Shuurman, J. G., de Croock, M. B. M., & Paas, F. G. W. C. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction*, *12*, 11-37.

Walker, N., & Fisk, A. D. (1995). To the gridiron. *Ergonomics in Design*, *3*, 7-13.

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory,* *and Cognition*, *19*(1), 203-222.

Wechsler, D. (1997). *Wechsler adult intelligence scale iii* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin and Review*, *9*, 185-211.